REMARKS

Claims 19-32 are pending. Claim 1 has been cancelled and new claims 19-32 drafted for ease of examination. Support for new claims 19-32 derives from the specification and claims as originally filed. For example, support for computational methods for the generation of primary libraries are described at page 7, line 22, through page 8 line 12, and page 10, line 9 through page 15, line 14; methods for the generation of secondary libraries from primary libraries are described at page 26, line 27, through page 30, line 27; methods for the generation of tertiary libraries are described at page 34, line 18, through page 19; page 40, line 14; support for the synthesis of variant proteins, beginning with the corresponding oligonucleotide sequences using multiple PCR can be found at pages 31-32; methods for isolating, purifying, and expressing the oligonucleotide sequences as proteins are well known in the art, and are described at pages 41-47 and in the Examples; and, the use of a computer workstation comprising a microprocessor is described at page 64, lines 1-6. Accordingly, the amendments do not present new matter and entry is proper.

Applicants thank the Examiner for withdrawing the rejection under 35 USC §112, second paragraph and the Double Patenting rejection.

Rejections under 35 U.S.C. § 101

Claim 1 is rejected under 35 U.S.C. § 101 for lacking a specific asserted utility or a well established utility. In rejecting claim 1, the Examiner's position appears to be that the claimed method generates a secondary library of undefined structure that is so general as to lack a real-world utility. The rejection is moot as applied to cancelled claim 1. Applicants respectfully submit that this rejection does not apply to newly added claims 19-32 for the following reasons.

Newly added claims 19-32 disclose the following inventions: (1) computational methods for generating a secondary library of protein variants (independent claim 19 and dependent claims 20-21) and methods for generating a tertiary library of protein variants (independent claim 22 and dependent claims 23-29); and, (2) an application of a computer program product (independent claim 30 and dependent claims 31-32).

The Examiner's basic position appears to be that there is no "specific and substantial utility", citing *In re Kirk*. As a preliminary matter, the Applicants first note that *In re Kirk* is a case dated prior to the new Utility Guidelines, and secondly that *Kirk* is directed to compositions of a new chemical class with a sole utility of "useful biological properties".

As to the first point, the Applicants respectfully draw the Examiner's attention to the Utility Guidelines:

In most cases, an applicant's assertion of utility creates a presumption of utility that will be sufficient to satisfy the utility requirement of 35 U.S.C. § 101. As the CCPA stated in In re Langer:

> As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope.

Thus, Langer and subsequent cases direct the Patent Office to presume that a statement of utility made by an applicant is true. For obvious reasons of efficiency and in deference to an applicant's understanding of his or her invention, when a statement of utility is evaluated, Patent Office personnel should not begin an inquiry by questioning the truth of the statement of utility. Instead, any inquiry must start by asking if there is any reason to question the truth of the statement of utility. This can be done by evaluating the logic of the statements made, taking into consideration any evidence cited by the applicant. If the asserted utility is credible (i.e., believable based on the record or the nature of the invention), a rejection based on "lack of utility" is not appropriate. Thus, Patent Office personnel should not begin an evaluation of utility by assuming that an asserted utility is likely to be false, based on the technical field of the invention or for other general reasons.

Compliance with § 101 is a question of fact. Thus, to overcome the presumption of truth that an assertion of utility by the applicant enjoys, Patent Office personnel must establish that it is more likely than not that one of ordinary skill in the art would doubt (i.e., "question") the truth of the statement of utility. To do this, Patent Office personnel must provide evidence sufficient to show that a person of ordinary skill in the art would consider the statement of asserted utility "false". A person of ordinary skill must have the benefit of both facts and reasoning in order to assess the truth of a statement. This means that if the applicant has presented facts that support

the reasoning used in asserting a utility, Patent Office personnel must present countervailing facts and reasoning sufficient to establish that a person of ordinary skill would not believe the applicant's assertion of utility (MPEP §2107.02IIIA). The initial evidentiary standard used during evaluation of this question is a preponderance of the evidence (i.e., the totality of facts and reasoning suggest that it is more likely than not that the statement of the applicant is false). It is respectfully submitted that the Examiner has not met this burden.

The claims are directed to specific methods of computationally generating libraries of proteins. As has been argued previously, these methods have a "real world" utility, as evidenced in several ways. First of all, a "real world" use has been shown because methods of protein design related to those of the present invention have been shown to work as claimed. See also U.S. Patent Nos. 6,188,965; 6,296,312; 6,403,312; 6,708,120; 6,792,356; PCT/US98/07254 and PCT/US01/40091. Such methods have been used to generate novel proteins with enhanced properties, see for example, U.S. Patent Nos. 6,682,923; 6,627,186; 6,514,729; and 6,746,853. See also, Steed et al, Science (2003), 301: 1895-1898, a copy of which is enclosed as Exhibit A; Hayes et al., PNAS, 99 (25): 15926-15931, a copy of which is enclosed as Exhibit B; and Luo et al., Protein Science (2002), 11: 1218-1226, a copy of which is enclosed as Exhibit C. Applicant also notes that the methodology described in these patents and scientific publications is not limited to enzymes, but applies to therapeutic proteins as well as any other type of proteins.

In further support of utility, the utility of these methods are recognized by those of skill in the art as useful techniques. A number of third parties have recognized the value of these methods. For example, in the article "Proteins from Scratch" (DeGrado, Science (1997), 278:80-81, a copy of which is enclosed as Exhibit D), biochemistry professor William F. DeGrado of the University of Pennsylvania School of Medicine, a world-renowned expert in protein structure, folding and design, comments on the computational platform designed by Dahiyat and Mayo in Science (1997), 278:82-87. This platform is an earlier version of the computational platform that has evolved and is claimed herein. Dr. DeGrado states:

> Not long ago, it seemed inconceivable that proteins could be designed from scratch. Because each protein sequence has an astronomical number of potential confirmations, it appears that only an experimentalist with the evolutionary life span of Mother Nature could design a sequence capable of folding into a single, well-defined three dimensional structure. But now on page 82 of

> this issue, Dahiyat and Mayo describe a new approach that makes de novo protein design as easy as running a computer.

Dr. DeGrado further states (col 1, paragraph 3):

> Thus, the problem of de novo protein design reduced to two steps: selecting a desired tertiary structure and finding a sequence that would stabilize this fold. Dahiyat and Mayo have now mastered the second step with spectacular success. They have distilled the rules, insights and paradigms gleaned from two decades of experiments into a single computational algorithm...Thus the rules of ...computational methods for de novo design may now be sufficiently defined to allow the engineering of a variety of proteins.

Thus, as can be seen from the selections cited above in Dr. DeGrado's article, Dr. DeGrado is commenting on the usefulness of the general method. Thus, Applicants respectfully dispute the Examiner's statement that "the article by DeGrado relates to the specific compound, Zinc finger protein". Dr. DeGrado is specifically discussing the computational design of Mayo and Dahiyat, not just a zinc finger protein.

Furthermore, Applicants respectfully submit that the Examiner has misunderstood DeGrado by his reference to the quotation at page 80 citing "de novo design is best approached by simultaneously considering all of the side chains in the protein-unfortunately, a very high order combinatorial problem". It is this very paragraph that goes on to discuss Dayhiyat and Mayo's DEE theorem to "efficiently search through sequence and side chain rotamer space" (see column 3, page 80, last sentence of second full paragraph). Thus the DeGrado article articulates that the Dahiyat and Mayo solution, which forms the basis of the present claims, is in fact very useful in the field of combinatorial evaluation.

Further, in 2002, Dr. Jeffery G. Saven, a well-known expert in protein design, has recently published a review of the state of the art in combinatorial protein libraries (see, Saven, JG, Curr. Op. Struct. Biol. (2002), 12:453-458, a copy of which is enclosed as Exhibit E, where he states at page 456, col. 1, 3rd paragraph, lines 6 – 13:

> Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but with combinatorial experiments, an enormous

> diversity of sequences can be generated at the control of the researcher.

Saven also states that

> Thus, methods for winnowing and focusing sequence space are a viral component of combinatorial protein design (see page 453, column 1, first paragraph) . . . Combinatorial methods are powerful tools for cases in which we have an incomplete understanding of molecular properties.

The Saven publication, while not prior art in the instant application, shows that it is known in the art that combinatorial library generation has "real world use". Thus, the discussions above regarding examples of actual utility by Applicant, as well as recognition to those skilled in the art of protein design and combinatorial library generation, meets the utility requirement under 35 USC § 101.

With respect to the follow on statement by Examiner, "or human growth hormone by Filikov", it is respectfully submitted that the human growth hormone was designed using the protein design automation computer program described in the recited claims. The other cited improved protein publications and patents were designed using the protein design automation program defined in the claims. These are examples of the breadth of the program. The uses are not limited to enzymes, but to any protein. In addition, the methods used to identify such improved proteins are used in this instant case.

The Examiner cites In re Kirk for the proposition that " We do not believe that it was the intention of the statutes to require the Patent Office, the courts or the public to play the sort of guessing game that might be involved if an applicant could satisfy the requirements of the statutes by indicating the usefulness of a claimed compound in terms of **possible use so general as to be meaningless**..." (emphasis by Examiner).

Applicant's initially point out that the *Kirk* case relates to a compound, not a method. Further, the methods recited in the present claims recite specifically defined steps that are understandable to those skilled in the art of computational biology and chemistry. The elements of the claims, including the use of scoring functions and probability distribution tables (claim 19), using PDA®, synthesizing and screening library sequences (claim 22) and using PDA® as well as a probability distribution table (claim 30) support a finding of utility as defined in 35 U.S.C. §101. **TT

In addition, the Applicants respectfully draw the Examiner's attention to the requirements as further outlined in the Guidelines:

> Where an applicant has specifically asserted that an invention has a particular utility, that assertion cannot simply be dismissed by Office personnel as being "wrong," even when there may be reason to believe that the assertion is not entirely accurate. Rather, Office personnel must determine if the assertion of utility is credible (i.e., whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided). An assertion is credible unless (a) the logic underlying the assertion is seriously flawed, or (b) the facts upon which the assertion is based are inconsistent with the logic underlying the assertion. Credibility as used in this context refers to the reliability of the statement based on the logic and facts that are offered by the applicant to support the assertion of utility.
>
> * * *
>
> . . . a prima facie showing [of no specific and substantial credible utility] must establish that it is more likely than not that a person of ordinary skill in the art would not consider that any utility asserted by the applicant would be specific and substantial.

Thus, the burden is shifted to the Examiner. The Applicants respectfully submit that this burden has not been met, and the rejection should be withdrawn.

The arguments made above with respect to 35 USC §101 are equally applicable to the rejection under 35 USC §112, first paragraph. The techniques described in the recited methods have a specific and well-established utility, and one skilled in the art would know how to use the claimed invention, particularly as demonstrated in the patents and scientific articles discussed above.

Lack of Utility under §112, 1st Paragraph

As argued above, there is sufficient utility under both 35 U.S.C. §§ 101 and 112 to meet the statutory requirements, and this rejection should be withdrawn.

Rejections under 35 U.S.C. § 112, first paragraph

Claim 1 is rejected under 35 U.S.C. 112, first paragraph for failing to comply with the written description requirement. In rejecting claim 1, the Examiner's position appears to be that the specification is enabling only for the design of enzymes (see page 5 of the final office

action). The rejection is moot as applied to cancelled claim 1. Applicants respectfully submit that this rejection does not apply to newly added claims 19-32 for the following reasons.

The Applicants acknowledge the Examiner's statement that the specification is enabling for methods utilizing enzymes, but respectfully disagree that other types of proteins are not enabled.

As stated infra regarding utility, Applicant respectfully submits that the application is enabled by the examples where a molecule whose coordinates were input into a computer, heavy side chain atoms were selected within a 4 Angstrom sphere around four catalytic residues. These heavy side chain atoms defined the variable residue positions for which a primary library was calculated. A probability table (Table 3) was calculated from the top 1000 sequences in the list (again see Table 3). Table 3 shows the number of occurrences of each of the amino acids selected for each position (i.e., 5 variable positions and 25 floated positions). One skilled in the art would readily be capable of extrapolating these examples to a variety of protein systems with a variety of functions, particularly when read in light of the specification (e.g. see Specification page 7, line 27 to page 9, line 5; page 34, line 22 to page 35, line 12). Thus these examples also show enablement.

With respect to the scope of the enabling disclosure not commensurate with the scope provided in the Specification, there is disclosure of using a computational design program, and preferably PDA® technology as embodiments of the invention. See Specification at page 2, lines 1-3; page 7, lines 9-12; and page 14, line 30 to page 15, line 5. In addition, the examples provide further enabling disclosure to one skilled in the art to practice this invention. As stated previously, the methodology is not limited to a particular kind of protein, and one skilled in the art would not be led to believe that this method is limited to enzymes. The method of the present invention is not limited to enzymes, since the modifications may be done to any proteins, not just enzymes. The methodology has been successfully employed in many non-enzyme proteins, e.g., TNF, GCSF, Interferon, etc. The publications cited in the section addressing the 35 USC §101 show the diversity of proteins that may be used. In addition, the article by Dr. Saven shows that those skilled in the art do not limit proteins by type (such as enzymes). The methodologies apply to any type of protein. The methodology requires that coordinates of a target protein be input.

There is nothing in the methodology that so limits it only to enzymes, and while the examples show enzyme modifications, these examples are just that, examples of how the technology works. The specification provides support for the use of any protein that may be used in this method. One skilled in the art would understand that this method may be used on any protein and not just limited to enzymes.

The Examiner cites the DeGrado reference at page 80 (See Office Action, page 6, first full paragraph). Applicants have two points; first of all, the applicants are not designing a protein de novo, which is the subject of the DeGrado quote, but are inputting the coordinates of a target protein. Inputting the coordinates of a target protein is the equivalent to enabling the analysis of that particular protein structure. The methodology employs known physio-chemical parameters of proteins, amino acids and rotamers to modify the target protein. Secondly, DeGrado also actually discusses the fact that this "very high combinatorial problem" is addressed by the Dahiyat and Mayo technique. Thus, DeGrado also supports a finding of enablement of the present techniques.

Thus for every protein (not just enzymes), the same methodology as recited in the instant claims is used.

There is no undue experimentation since the specification enables one skilled in the art to practice the invention using the specifically recited steps in the claims. The Examiner refers to Cys, Pro and Gly not being used in an Example in the specification. Applicants' respectfully refer the Examiner to page 17, lines 30-35, where the specification discloses the basis behind using, or not using certain amino acids in certain situations. To one skilled in the art of protein design, this is not undue experimentation but a design choice. With respect to the Examiner's comments regarding SO2 and water being removed, Applicants' respectfully refer the Examiner to page 15, lines 6-25 for the discussion on backbone structure preparation, as well as the discussion on backbone preparation above.

Applicants respectfully point to In re Goffe, 191 USPQ429 (CCPA 1976), where the court stated:

> For all practical purposes, the Board would limit Appellant to claims involving the specific materials disclosed in the examples,

> so that a competitor seeking to avoid infringing the claims would merely have to follow the disclosure in the subsequently issued patent to find a substitute. However, to provide effective incentives, claims must adequately protect inventors. To demand that the first to disclose shall limit his claims to what he has found to work or to materials which meet the guidelines specified for "preferred" materials in a process such as the one herein involved would not serve the constitutional propose of promoting progress in the useful arts.

Additionally, in In re Angstadt, 190 USPQ 214, 218 (CCPA 1976), the court further stated:

> Appellants have apparently not disclosed every catalyst which will work; they have apparently not disclosed every catalyst which will not work. The question, then, is whether in an unpredictable art, section 112 requires disclosure of a test with every species covered by a claim. To require such a complete disclosure would apparently necessitate a patent application or applications with "thousands" of examples or the disclosure of "thousands" of catalysts along with information as to whether each exhibits catalytic behavior resulting in the production of hydroperoxides. More importantly, such a requirement would force an inventor seeking adequate patent protection to carry out a prohibitive number of actual experiments. This would tend to discourage inventors from filing patent applications in an unpredictable area since the patent claims would have to be limited to those embodiments which are expressly disclosed.

Therefore, in conclusion, Applicants submit that the Specification taken in conjunction with the state of the art at the time the invention was filed fully enables a person skilled in the art to practice the method of the invention without undue experimentation. Applicants respectfully request reconsideration and withdrawal of the rejection.

Applicants respectfully submit that the specification enables a method for computationally generating a genus of secondary libraries comprising variant sequences in which the starting protein structure (*i.e.* target protein or scaffold protein) can be any protein for which a three dimensional structure is known or can be generated. In addressing the written description requirement under 35 U.S.C. § 112, the Federal Circuit in University of *California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997), stated:

> A description of a genus of cDNAs may be achieved by means of a recitation of a representative number of cDNAs, defined by nucleotide sequence, falling with the scope of the genus or of a recitation of structural features common to the members of the genus, which features constitute a substantial portion of the genus. This is analogous to enablement of a genus under Section 112, para. 1, by showing the enablement of a representative number of species within the genus. *See Angstadt*, 537 F.2d at 502-03 (deciding that applicants "are *not* required to disclose *every* species encompassed by their claims their claims even in an unpredictable art and that the disclosure of forty working examples sufficiently described subject matter of claims directed to a generic process) . . . See also *In re Grimme*, 274 F.2d, 949, 952 ("[I]t has been consistently held that the naming of one member of such a group is not, in itself, a proper basis for a claims to the entire group. However, it may not be necessary to enumerate a plurality of species if a genus is sufficiently identified in an application by other appropriate language.").

In support of the position that Applicants' have designed many proteins that are not "enzymes", Applicants enclose herewith a number of publications that are both prior and subsequent to the filing date of the present application. These are not offered to augment the disclosure of the application; rather, the work is presented to show that present invention is enabled for any protein for which a defined set of coordinates can be generated. *See In re Wilson*, 135 USPQ 442, 444 (CCPA 1962); *Ex parte Obukowicz*, 27 USPQ 2d 1063 (BPAI 1993); *Gould v. Quigg*, 3 USPQ 2d 1302,1305 (Fed. Cir. 1987):

> "it is true that a later dated publication cannot supplement an insufficient disclosure in a prior dated application to render it enabling. In this case the later dated publication was not offered as evidence for this purpose. Rather, it was offered . . . as evidence that the disclosed device would have been operative" printed publications.

For example, the enclosed articles describe computationally designed GCSF (US 6627186 and Luo P et al., Protein Science 11, 1218-1226 (2002); enclosed herein as Exhibits A and B), Interferon Beta (US 6514729, enclosed herein as Exhibit C) and TNF-alpha (US publication No. 2003/138401 and Steed PM et al, Science 301, 1895-1898 (2003); enclosed herein as Exhibits D and E), for example. These non-enzymatic proteins have a variety of

structures and have all been successfully designed. Thus, it is improper to limit the scope of this invention to just "enzymes".

Applicant's respectfully point out the new claim 22 specifically recites PDA® in response to the Examiner's statement at page 6 first full paragraph of the Office Action. PDA is a preferred embodiment of Applicant's invention, but this particular computational approach is not necessarily required. Applicant's also respectfully dispute that the recited method claims are "the shot in the dark, genetic approach" (see Office Action, page 6, first full paragraph). The approach is rational, not random ("shot in the dark"). Applicant's again reiterate that one skilled in the art would be enabled to practice the steps of the method without undue experimentation.

The articles, patents and patent applications discussed above, support the enablement of the methods disclosed in the pending claims. Importantly, the methods apply to proteins in general, regardless of whether the protein is an enzyme, as described in the example, or an antibody, cell surface receptor, or other protein of interest.

Accordingly, Applicants respectfully submit that the specification fully enables the present claims, and respectfully request withdrawal of the rejection under 35 U.S.C. § 112, first paragraph.

Rejections under 35 U.S.C. § 102

Claim 1 is rejected under 35 U.S.C. § 102(b) as being anticipated by Fechteler et al., 1995, *J. Mol. Biol.*, 13: 114-31. In rejecting claim 1, the Examiner's position appears to be that at page 128, that designing a protein model is the same concept as generating a second variant library, and thus, Fechteler describes the same computational method as taught in the instant application. The rejection is moot as applied to cancelled claim 1. Applicants respectfully submit that this rejection does not apply to newly added claims 19-32 for the following reasons.

To anticipate a claim under 35 U.S.C. § 102(b), a reference must teach every element of the rejected claim (MPEP § 2131).

Applicants respectfully submit that Fechteler teaches a method for predicting the three-dimensional structure in insertion /deletion regions of a protein structure that combines cluster

analysis with a geometric scoring criteria. Fechteler uses clustering with geometric criteria to narrow the list of fragment options when attempting to fit structural fragments onto the existing template structure. The modeling in Fechteler always takes place for a single unique protein sequence (i.e. although structural variants are created, no sequence variants are created).

Applicants also respectfully reiterate that Fechteler does not teach synthesis of proteins of the invention – the methods sections on pages 128-129 of the Fechteler reference merely spell out the details of the Fechteler structure prediction method, and the only entity that can be produced by the method of Fechteler is a theoretical list of 3-D coordinates for placement of atoms in space. Indeed, as Fechteler was completely focused on predicting the structure of a protein based on its sequence, the method was only applied to sequences that had already been synthesized and characterized – that is, the order of application is reversed relative to the Applicants' method in which variant sequence libraries are designed and then produced.

Further, the Fechteler reference does not create novel variant sequences. All of the sequences identified are the same as in the initial set.

In contrast to Fechteler, claims 19-32 use physico-chemical scoring functions (e.g. van der Waals, hydrogen bonding, etc.), probability tables and protein design automation to computationally filter variant protein sequences and generate a primary list of variant proteins. The current invention then further generates a secondary library of variant protein sequences by combining a plurality of variant amino acid residues. There is also no discussion or teaching in Fechteler of combining a plurality of their database fragments. Fechteler does not teach or suggest the use of scoring functions, probability tables, protein design automation, or the design of variant protein sequences and libraries.

Hence, Fechteler does not anticipate the claimed subject matter. Withdrawal of the rejection under 35 U.S.C. § 102(b) is requested.

The Examiner is invited to contact the undersigned at (415) 781-1989 if any issues may be resolved in that manner.

Respectfully submitted,
DORSEY & WHITNEY LLP

Dated: 10/21/04

By: Robin M. Silva

Four Embarcadero Center
Suite 3400
San Francisco, California 94111-4187
Telephone: (415) 781-1989
Fax No. (415) 398-3249

Robin M. Silva, Reg. No. 38,304
Filed under 37 C.F.R. § 1.34(a)

estimated between 20,000 and 40,000 metric tons (*10*, *12*). Assuming that pre-Incan technologies operated at efficiencies comparable to *huayras* and recognizing the initial presence of grades as rich as 25% Ag (*11*), our data imply that several thousand tons of silver were produced in pre-Incan times. Although major new archaeological discoveries in the Andes remain a distinct possibility, the likelihood seems equally probable that most of this silver was recycled and transported elsewhere in the Americas before conquest, or eventually exported overseas by the Spanish.

References and Notes

1. H. Lechtman, in *The Coming of the Age of Iron*, P. J. Wertime, J. D. Muhly, Eds. (Yale Univ., New Haven, CT, 1980), pp. 267–334.
2. R. L. Burger, R. B. Gordon, *Science* 282, 1108 (1998).
3. E. P. Benson, Ed., *Pre-Columbian Metallurgy of South America* (Dumbarton Oaks, Washington, DC, 1979).
4. I. Shimada, S. Epstein, A. K. Craig, *Science* 216, 952 (1982).
5. H. Lechtman, in *Tiwanaku and Its Hinterland: Archaeology and Paleoecology of an Andean Civilization Vol. 2*, A. L. Kolata, Ed. (Smithsonian Institution, Washington, DC, 2002), pp. 404–434.
6. K. O. Bruhns, *Ancient South America* (Cambridge Univ., Cambridge, 1994).
7. W. E. Rudolph, *Am. Geogr. Soc.* 26, 529 (1936).
8. M. Vuille, *Int. J. Climatol.* 19, 1579 (1999).
9. Materials and methods are available as supporting material on *Science* Online.
10. P. J. Bartos, *Econ. Geol.* 95, 645 (2000).
11. W. E. Wilson, A. Petrov, *Miner. Rec.* 30, 9 (1999).
12. P. J. Bakewell, *Miners of the Red Mountain: Indian Labor in Potosí 1545–1650* (Univ. of New Mexico, Albuquerque, NM, 1984).
13. R. Peele, *Sch. Mines Q.* 15, 8 (1893).
14. L. Renberg, I. M. Wik-Persson, O. Emteryd, *Nature* 368, 323 (1994).
15. M. L. Brännvall, R. Bindler, O. Emteryd, I. Renberg, *J. Paleolimnol.* 25, 421 (2001).
16. C. Gobeil, *Environ. Sci. Technol.* 33, 2953 (1999).
17. L. Rivera-Duarte, A. R. Flegal, *Geochim. Cosmochim. Acta* 58, 3307 (1994).
18. B. B. Wolfe, *Palaeogeogr. Palaeoclim. Palaeoecol.* 176, 177 (2001).
19. A. S. Ek, I. Renberg, *J. Paleolimnol.* 26, 89 (2001).
20. A. F. Bandelier, *The Islands of Titicaca and Koati* (Hispanic Society of America, New York, 1910).
21. P. R. Williams, *World Archaeol.* 33, 361 (2002).
22. A. L. Kolata, *The Tiwanaku: Portrait of an Andean Civilization* (Blackwell, Oxford, UK, 1993).
23. M. B. Abbott, M. W. Binford, M. Brenner, K. R. Kelts, *Quat. Res.* 47, 169 (1997).
24. B. S. Bauer, C. Stanish, *Ritual and Pilgrimage in the Ancient Andes: The Islands of the Sun and the Moon* (Univ. of Texas, Austin, TX, 2001).
25. L. G. Thompson, E. Mosley-Thompson, J. F. Bolzan, B. R. Koci, *Science* 229, 971 (1985).
26. A. K. Craig, in *Precious Metals, Coinage, and the Changes in Monetary Structures in Latin America, Europe, and Asia*, E. van Cauwenberghe, Ed. (Leuven Univ. Press, Leuven, Netherlands, 1989), pp. 159–183.
27. G. E. Ericksen, R. G. Luedke, R. L. Smith, R. P. Koeppen, F. Urquidi, *Episodes* 13, 5 (1990).

28. Supported by the U.S. NSF-ESH (M.B.A), Natural Sciences and Engineering Research Council of Canada (A.P.W.) and Geological Society of America. We are grateful to H. Lechtman, M. Bermann, C. Cooke, and journal reviewers for their insightful comments; G. Seltzer for assistance in the field; and S. Root for support in the laboratory.

Supporting Online Material
www.sciencemag.org/cgi/content/full/301/5641/1893/DC1
Materials and Methods
Fig. S1
Tables S1 and S2

9 June 2003; accepted 18 August 2003

Inactivation of TNF Signaling by Rationally Designed Dominant-Negative TNF Variants

Paul M. Steed,* Malú G. Tansey,*† Jonathan Zalevsky,* Eugene A. Zhukovsky, John R. Desjarlais, David E. Szymkowski, Christina Abbott, David Carmichael, Cheryl Chan, Lisa Cherry, Peter Cheung, Arthur J. Chirino, Hyo H. Chung, Stephen K. Doberstein, Araz Eivazi, Anton V. Filikov, Sarah X. Gao, René S. Hubert, Marian Hwang, Linus Hyun, Sandhya Kashi, Alice Kim, Esther Kim, James Kung, Sabrina P. Martinez,† Umesh S. Muchhal, Duc-Hanh T. Nguyen, Christopher O'Brien, Donald O'Keefe, Karen Singer, Omid Vafa, Jost Vielmetter, Sean C. Yoder, Bassil I. Dahiyat‡

Tumor necrosis factor (TNF) is a key regulator of inflammatory responses and has been implicated in many pathological conditions. We used structure-based design to engineer variant TNF proteins that rapidly form heterotrimers with native TNF to give complexes that neither bind to nor stimulate signaling through TNF receptors. Thus, TNF is inactivated by sequestration. Dominant-negative TNFs represent a possible approach to anti-inflammatory biotherapeutics, and experiments in animal models show that the strategy can attenuate TNF-mediated pathology. Similar rational design could be used to engineer inhibitors of additional TNF superfamily cytokines as well as other multimeric ligands.

TNF is a proinflammatory cytokine that can complex two TNF receptors, TNFR1 (p55) and TNFR2 (p75), to activate signaling cascades controlling apoptosis, inflammation, cell proliferation, and the immune response (*1*–*5*). The 26-kD type II transmembrane TNF precursor protein, expressed on many cell types, is proteolytically converted into a soluble 52-kD homotrimer (*6*). An elevated serum level of TNF is associated with the pathophysiology of rheumatoid arthritis (RA), inflammatory bowel disease, and ankylosing spondylitis (*1*, *7*, *8*), and molecules that inhibit TNF signaling have demonstrated clinical efficacy in treating some of these diseases (*9*, *10*).

We have engineered dominant-negative TNF (DN-TNF) variants that inactivate the native homotrimer by a sequestration mechanism that blocks TNF bioactivity (fig. S1). Protein design automation (PDA), an in silico method that predicts protein variants with improved biological properties (*11*–*13*), was used to introduce single or double amino acid changes into TNF (Fig. 1A) to generate the desired biological profile while maintaining the overall structural integrity of the molecule. Specifically, our goal was to design homotrimeric TNF variants that (i) have decreased receptor binding, (ii) sequester native TNF homotrimers from TNF receptors by formation of inactive native:variant heterotrimers, (iii) abolish TNF signaling in relevant biological assays, and (iv) are easily expressed and purified in large quantities from bacteria. Variants were tested for TNF receptor activation in cell-based assays, and nonagonistic variants were then checked for their ability to antagonize native TNF in cell and animal models. Subsequently, we evaluated assembly state, receptor binding, and heterotrimer formation for several variants.

The computational design strategy used crystal structures of native and variant TNF trimers as templates for the simulations. Analysis of a homology model of the TNF-receptor complex revealed several distinct regions of the cytokine that make multiple direct contacts with its receptors (Fig. 1A), including interfaces rich in hydrophobic and electrostatic interactions. We ran simulations to select nonimmunogenic point mutations that would disrupt receptor interactions while preserving the structural integrity of the TNF variants and their ability to assemble into heterotrimers with native TNF (*14*). Many of the designed TNF variants displayed markedly reduced binding to TNFR1 and TNFR2, and several combinations of potent single mutations further decreased binding (Fig. 1B and fig. S2). As predicted by analysis of the TNF-TNFR structural complex, combinations of the most potent single mutations at different interaction domains (e.g., A145R and

Xencor, 111 West Lemon Avenue, Monrovia, CA 91016, USA.

*These authors contributed equally to this work.
†Present address: Department of Physiology, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390, USA.
‡To whom correspondence should be addressed. E-mail: baz@xencor.com

I97T) were frequently additive or synergistic. Moreover, our data extend results of previous studies (*15*, *16*) in demonstrating that certain substitutions can alter the specificity of receptor interactions; for example, I97T and A145R/I97T show greater relative binding to TNFR2 than to TNFR1 (Fig. 1B and fig. S2).

Homotrimers of several designed TNF variants exhibited >10,000-fold reduction in their ability to activate two major signaling pathways downstream of TNF receptor activation. Specifically, single variants such as Y87H and A145R and the double variant A145R/Y87H were unable to bind to either TNFR1 or TNFR2 in cell-free assays (Fig. 1B and fig. S2) and failed to activate either caspase (Fig. 1C) or nuclear factor κB (NF-κB)–mediated luciferase expression (Fig. 1D) relative to native TNF. In contrast, those variants that still bound TNFRs also activated TNF signaling; in some cases (e.g., F144N) more potently than native TNF. Disruption of two receptor interfaces (e.g., A145R/Y87H or A145R/I97T) effectively destroyed the residual agonism detected with some single-point TNF variants. Furthermore, the importance of using multiple screening criteria to evaluate DN-TNF bioactivity was revealed by variants such as A145R/I97T, which had virtually no TNF-like activity in either cell-based assay yet displayed appreciable TNFR2 binding affinity.

We subsequently tested nonagonistic TNF variants for their ability to act as dominant-negative inhibitors by measuring their capacity to block native TNF activity in cell-based assays. We evaluated dose-dependent TNF antagonism (*14*) by mixing increasing concentrations of variants with native TNF (5 ng/ml) for 1.5 hours and measuring the caspase activity induced by these mixtures after addition to U937 cells (Fig. 2A). At concentrations as low as twofold that of native TNF (10 ng/ml), A145R/Y87H and

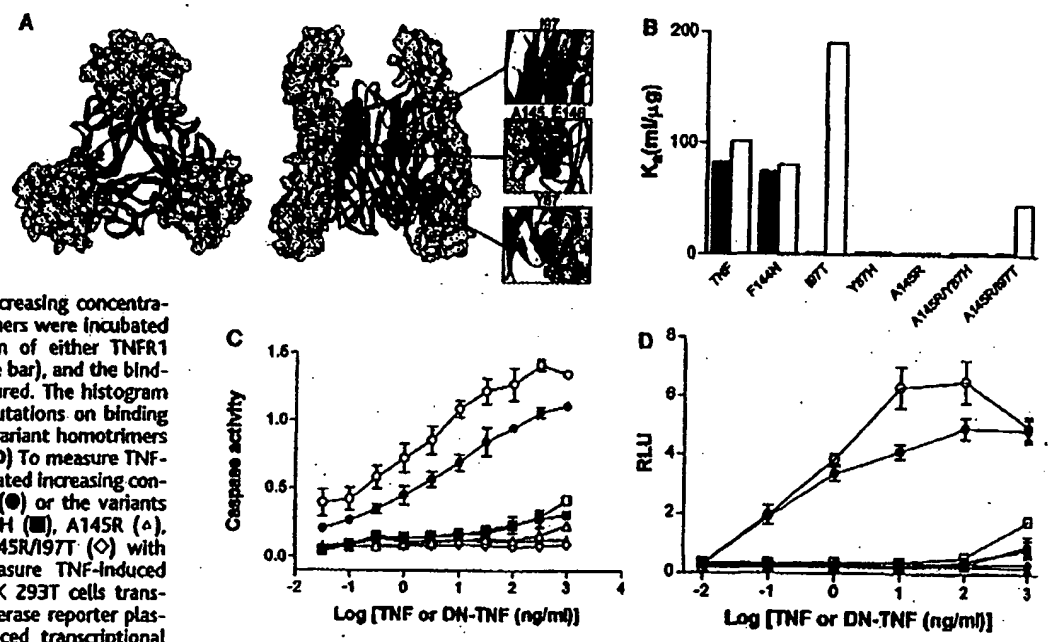


Fig. 1. DN-TNF variants have impaired TNF receptor binding and signaling. (**A**) Structural schematic of human TNF trimer–TNFR1 complex with major contacts between ligand and receptor highlighted by solid surfaces (green). Locations of representative mutated residues substituted in dominant-negative variants are shown in boxes. (**B**) Increasing concentrations of DN-TNF homotrimers were incubated with a fixed concentration of either TNFR1 (black bar) or TNFR2 (white bar), and the binding affinity (K_a) was measured. The histogram illustrates the effect of mutations on binding affinity between DN-TNF variant homotrimers and TNF receptors. (**C** and **D**) To measure TNF-induced signaling, we incubated increasing concentrations of native TNF (●) or the variants F144N (○), I97T (□), Y87H (■), A145R (△), A145R/Y87H (▲), and A145R/I97T (◇) with either U937 cells to measure TNF-induced caspase activity (C) or HEK 293T cells transfected with an NF-κB–luciferase reporter plasmid to measure TNF-induced transcriptional activation (D). DN-TNF variants, especially the double mutants, have reduced TNF receptor binding and signaling activity. RLU, relative luciferase units; Caspase activity, arbitrary units normalized to V_{max}.

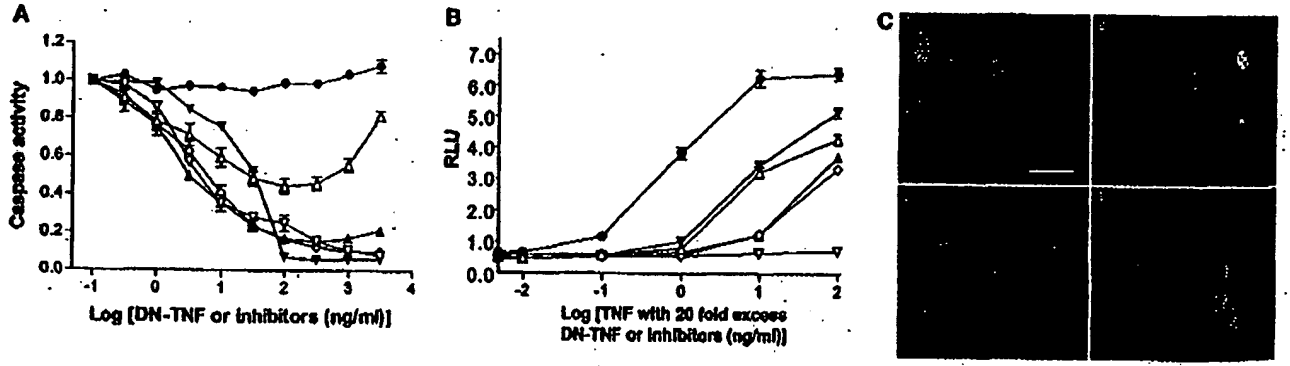


Fig. 2. DN-TNF variants inhibit TNF-mediated intracellular signaling. (**A**) Inhibition of caspase activation. Native TNF (5 ng/ml) was mixed with buffer (●) or with increasing concentrations of either TNF variants A145R (△), A145R/Y87H (▲), or A145R/I97T (◇), the soluble Fc-TNFR2 fusion etanercept (▽), or the TNF monoclonal antibody infliximab (▼). After 1.5 hours of incubation in exchange buffer (*14*), these mixtures were applied to U937 cells to stimulate caspase activity. (**B**) Inhibition of NFκB pathway activation. Native TNF (●) (25 μg/ml) was mixed with 20-fold excess (by mass) of A145R (△), A145R/Y87H (▲), A145R/I97T (◇), etanercept (▽), or infliximab (▼). These mixtures were serially diluted and applied to HEK 293T cells for 12 hours to induce NF-κB–luciferase reporter activity. (**C**) A145R/Y87H inhibits native TNF-induced nuclear translocation of the p65-RelA subunit of NF-κB. Immunofluorescence studies show subcellular localization of NF-κB in HeLa cells treated with buffer (panel 1), native TNF (10 ng/ml) (panel 2), A145R/Y87H (100 ng/ml) (panel 3), or the combination of native TNF and variant A145R/Y87H (panel 4). RLU, relative luciferase units; Caspase activity, arbitrary units normalized to V_{max}. Scale bar, 25 μm.

A145R/I97T attenuated TNF-induced caspase activity by 50%, and at 20-fold excess, activity was reduced to baseline. The in vitro potency (by mass) of these variants is comparable to that of a soluble Fc-TNFR2 fusion (etanercept) and more potent than that of an antibody to TNF (infliximab), two marketed anti-TNF therapies, supporting the potential utility of this mechanism. Similarly, at 20-fold excess over native TNF, single-point (A145R, I97T, Y87H) and particularly double-point (A145R/Y87H, A145R/I97T) variants decreased caspase activation (fig. S3) as well as TNF-induced transcriptional activation by NF-κB in human embryonic kidney (HEK) 293T cells (Fig. 2B). Consistent with these results, the TNF variant A145R/Y87H (at 10-fold excess over native TNF) blocked TNF-induced nuclear translocation of the NF-κB p65-RelA subunit in HeLa cells (Fig. 2C). Thus, a number of variants neutralized TNF-induced caspase and NF-κB–mediated transcriptional activity over a wide range of native TNF concentrations, including the clinically relevant range of 100 to 200 pg/ml found in the synovial fluid of RA patients (*17–19*).

To demonstrate that the mechanism of TNF inhibition requires the formation of heterotrimeric complexes with native TNF, we measured the relation between heterotrimer levels and inhibition of TNF-induced signaling (*14*). We generated heterotrimeric complexes by mixing a fixed amount of FLAG-tagged native TNF with increasing concentrations of His-tagged TNF variants. A part of this material was used in a sandwich enzyme-linked immunosorbent assay (ELISA) (Fig. 3A, open symbols) to detect the formation of His-FLAG heterotrimers, and the remainder was applied to U937 cells to detect TNF-mediated caspase activation (Fig. 3A, closed symbols). The extent of heterotrimer formation of A145R/Y87H or A145R/I97T with native TNF correlated with a decrease in caspase activation, demonstrating an inverse relation between signaling and heterotrimer formation. As expected, etanercept activity is independent of TNF monomer exchange (Fig. 3A, open circles) because etanercept binds to the TNF trimer. To directly visualize heterotrimer formation, we mixed FLAG-tagged native TNF with His-tagged DN-TNF and resolved the exchanged products using native polyacrylamide gel electrophoresis (PAGE) (Fig. 3B) (*20*). Electrophoresis of equimolar quantities of mixed DN-TNF and native TNF resolved the variant homotrimer, 1:2 and 2:1 native:variant heterotrimers, and native homotrimer in approximately the expected 1:3:3:1 ratio (Fig. 3B, lane 10:10). Western blot analyses (*14*) with antibodies against the epitope tags confirmed the composition of the intermediate species (fig. S4). Stochastic equilibrium modeling of native and variant TNF heterotrimer assembly predicts that 10-fold excess of variant homotrimer causes the loss of more than 99% of homotrimeric native TNF, primarily into 1:2 native:variant heterotrimers, and our results confirmed this (Fig. 3B, lane 10:100). Exchange reactions between native and variant TNF reached ~80% completion at 20 min, and essentially all the native homotrimer was depleted after 90 min (fig. S5). Finally, we confirmed that biological activity of variants requires exchange into heterotrimeric complexes with native TNF. Specifically, our most potent variants (e.g., A145R/Y87H) failed to block caspase activity induced by chemically cross-linked native TNF homotrimers (*14*), which are unable to dissociate to allow exchange with variant TNFs (fig. S6).

The most potent in vitro inhibitors were selected for testing in vivo, to further study the mechanism and to begin therapeutic lead candidate identification. We tested the bioactivity of variant homotrimer and native:variant heterotrimers in the D-galactosamine (GalN)-sensitized mouse model, which demonstrated that DN-TNF homotrimers, and heterotrimers with native TNF, are devoid of agonist activity and efficiently exchange with endogenous TNF in vivo. GalN is a known specific hepato-toxin that can increase the sensitivity of mice to human TNF by 1000-fold (*21*, *22*). Native human TNF (30 μg/kg) induced severe hepatocellular apoptosis and lethality, consistent with previous reports

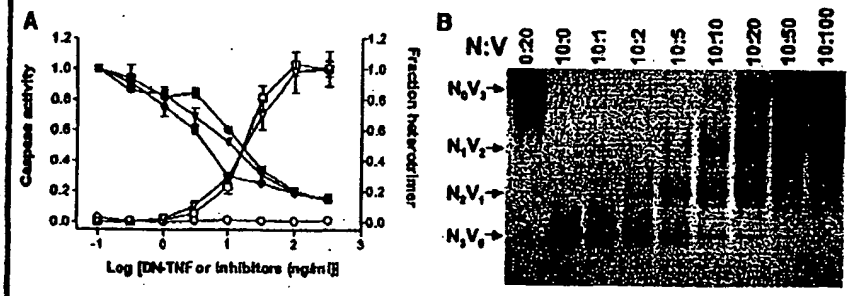


Fig. 3. DN-TNF variants inhibit signaling by sequestering native TNF into inactive heterotrimers. (A) Inverse correlation between heterotrimer formation and caspase activity. Native TNF was mixed in exchange buffer (*14*) with A145R/Y87H (▽, ▼), A145R/I97T (□, ■), or etanercept (○, ●), as described for Fig. 2. A part of each mixture was analyzed by a sandwich ELISA to detect native:variant TNF (open symbols), and the remainder was used to stimulate caspase activity in U937 cells (closed symbols). Caspase activity, arbitrary units normalized to V_{max}. (B) Native gel analysis of heterotrimer formation with various ratios of native (N) and DN-TNF (V). FLAG-tagged native TNF was incubated alone (N_3:V_0, lane 10:0) or with increasing concentrations of His-tagged variant A145R/Y87H (lanes 10:1 to 10:100) before native gel electrophoresis to determine heterotrimer formation. The differences in isoelectic point conferred by the epitope tags allowed for resolution of all possible trimer species (N_3:V_0, N_2:V_1, N_1:V_2, and N_0:V_3). Increasing concentrations of DN-TNF variant caused the redistribution of native TNF into both heterotrimers, and at 10-fold excess all detectable native TNF was consumed.

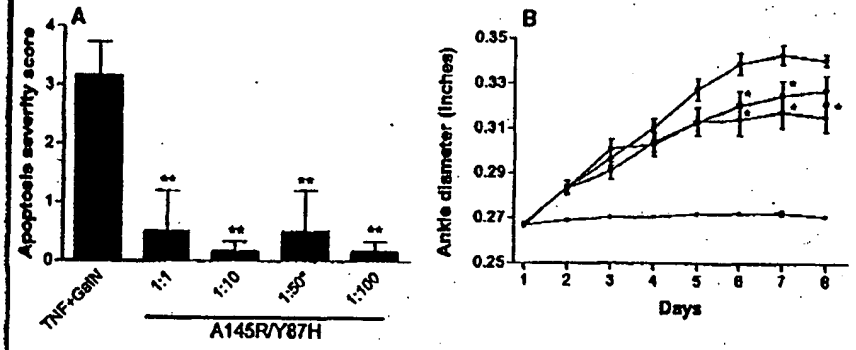


Fig. 4. DN-TNF variants exhibit efficacy in vivo. (A) Effect of heterotrimers of various ratios in the galactosamine-sensitized mouse model of human TNF-induced endotoxemia. Native human TNF was dosed at 30 μg/kg and A145R/Y87H was dosed at the indicated ratios to a fixed native human TNF dose of 30 μg/kg except at the 1:50* ratio, where A145R/Y87H was dosed in 50-fold excess of native human TNF (75 μg/kg). Livers were harvested and samples were blinded and scored for apoptotic damage on a scale of 0 to 4 as described (*14*). **$P < 0.05$. (B) Efficacy of A145R/Y87H in the rat 7-day established CIA model. A145R/Y87H was modified to introduce a PEG moiety at residue 31 of a non–epitope-tagged molecule as described (*14*). One group of four animals was nonarthritic (●); the remaining animals were collagen treated and, after the onset of symptoms, they were randomized into groups of eight. Animals were treated with vehicle (○), variant at 10 mg/kg twice daily dosing (□), or variant at 2 mg/kg subcutaneously with an intravenous loading dose of 2 mg/kg on the first treatment day (■). Measurements of ankle diameter were made daily by caliper. *$P < 0.05$.

(*23*); in contrast, A145R/Y87H dosed as high as 30 mg per kilogram of mouse body weight (along with native TNF at 30 μg/kg) resulted in no mortality or hepatotoxicity (Fig. 4A) (*24*). Similarly, lethal doses of native human TNF (30 μg/kg) mixed before injection with varying ratios of A145R/Y87H produced no TNF-induced damage. This protection was observed at native: variant ratios as low as 1:1 and with a superlethal dose of TNF (Fig. 4A). Further, sandwich ELISA analyses of serum samples indicated that a substantial portion (30 ng/ml) of administered A145R (3 mg/kg) was in heterotrimers with the endogenous mouse TNF at 1 hour.

A145R/Y87H was next assessed in a model of chronic disease as an initial test of the DN-TNF antagonism mechanism in a disease-relevant setting. We selected the rat 7-day established collagen-induced arthritis (CIA) model because it simulates chronic autoimmune joint disease and can be treated by TNF blockade (*25*). When dosed after the onset of symptoms, only interventions with rapid onset of action would be able to affect disease progression in this model, thus requiring rapid exchange in vivo of TNF variants with endogenous TNF. To ensure that there were no confounding in vivo effects of using affinity-tagged variants, we produced A145R/Y87H that lacked such tags. Further, to decrease in vivo clearance, we added one polyethylene glycol (PEG; ~5 kD/molecule) to each monomeric subunit of A145R/Y87H. This modification had no effect on the dominant-negative properties of the molecule in vitro (fig. S7). A145R/Y87H reduced joint swelling in the CIA model when dosed once daily at 2.0 mg/kg subcutaneously with a loading dose of 2.0 mg/kg and twice daily at 10 mg/kg intravenously (Fig. 4B). These results demonstrate the potential of DN-TNFs to inhibit TNF-mediated inflammation and verify that exchange occurs rapidly enough to affect progression of acute symptoms when dosed therapeutically.

Given their high-yield bacterial production, theoretical low immunogenicity, and unique mechanism of action, DN-TNFs show potential as a new class of anti-inflammatory therapy, particularly because existing methodologies (i.e., PEG modification) can be used to further enhance their pharmacokinetic properties (*26*, *27*). Further, we propose that this dominant-negative approach should be tested for its potential to create inhibitors of other multimeric extracellular signaling molecules, in particular other members of the TNF superfamily (e.g., RANKL, CD40L, and BAFF) that have been implicated in human pathophysiology (*28*, *29*).

References and Notes

1. B. B. Aggarwal, A. Samanta, M. Feldmann, in *Cytokine Reference*, J. J. Oppenheim, M. Feldmann, Eds. (Academic Press, London, 2000).
2. G. Chen, D. V. Goeddel, *Science* 296, 1634 (2002).
3. D. J. MacEwan, *Cell Signal.* 14, 477 (2002).
4. M. P. Boldin, T. M. Goncharov, Y. V. Goltsev, D. Wallach, *Cell* 85, 803 (1996).
5. G. M. Cohen, *Biochem. J.* 326, 1 (1997).
6. S. R. Ruuls *et al.*, *Immunity* 15, 533 (2001).
7. M. Feldmann, R. N. Maini, *Annu. Rev. Immunol.* 19, 163 (2001).
8. B. Beutler, *Immunity* 15, 5 (2001).
9. M. Feldmann, *Nature Rev. Immunol.* 2, 364 (2002).
10. R. Goldbach-Mansky, P. E. Lipsky, *Annu. Rev. Med.* 54, 197 (2003).
11. R. J. Hayes *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 99, 15926 (2002).
12. A. V. Filikov *et al.*, *Protein Sci.* 11, 1452 (2002).
13. P. Luo *et al.*, *Protein Sci.* 11, 1218 (2002).
14. Materials and methods are available as supporting material on *Science* Online.
15. J. Yamagishi *et al.*, *Protein Eng.* 3, 713 (1990).
16. X. M. Zhang, L. Weber, M. J. Chen, *J. Biol. Chem.* 267, 24069 (1992).
17. G. Steiner *et al.*, *Rheumatology* 38, 202 (1999).
18. T. Horiuchi *et al.*, *Endocr. J.* 46, 643 (1999).
19. A. K. Ulfgren *et al.*, *Arthritis Rheum.* 43, 2391 (2000).
20. P. Ameloot, W. Declercq, W. Fiers, P. Vandenabeele, P. Brouckaert, *J. Biol. Chem.* 276, 27098 (2001).
21. I. Hishinuma *et al.*, *Hepatology* 12, 1187 (1990).
22. S. Sakaguchi, S. Furusawa, K. Yokota, M. Takayanagi, Y. Takayanagi, *Int. J. Immunopharmacol.* 22, 935 (2000).
23. P. Brouckaert, C. Libert, B. Everaerdt, W. Fiers, *Lymphokine Cytokine Res.* 11, 193 (1992).
24. P. M. Staed *et al.*, data not shown.
25. A. M. Bendele *et al.*, *Arthritis Rheum.* 43, 2648 (2000).
26. K. Sreekrishna *et al.*, *Biochemistry* 28, 4117 (1989).
27. Y. P. Li *et al.*, *Biol. Pharm. Bull.* 24, 666 (2001).
28. C. F. Ware, *J. Exp. Med.* 192, F35-8 (2000).
29. R. M. Locksley, N. Killeen, M. J. Lenardo, *Cell* 104, 487 (2001).

30. We thank M. Ary for technical assistance with the manuscript.

Supporting Online Material
www.sciencemag.org/cgi/content/full/301/5641/1895/DC1
Materials and Methods
References
Figs. S1 to S7

9 December 2002; accepted 14 August 2003

The Dog Genome: Survey Sequencing and Comparative Analysis

Ewen F. Kirkness,[1] Vineet Bafna,[2*] Aaron L. Halpern,[2*] Samuel Levy,[2*] Karin Remington,[2*] Douglas B. Rusch,[2*] Arthur L. Delcher,[1] Mihai Pop,[1] Wei Wang,[1] Claire M. Fraser,[1] J. Craig Venter[2]

A survey of the dog genome sequence (6.22 million sequence reads; 1.5× coverage) demonstrates the power of sample sequencing for comparative analysis of mammalian genomes and the generation of species-specific resources. More than 650 million base pairs (>25%) of dog sequence align uniquely to the human genome, including fragments of putative orthologs for 18,473 of 24,567 annotated human genes. Mutation rates, conserved synteny, repeat content, and phylogeny can be compared among human, mouse, and dog. A variety of polymorphic elements are identified that will be valuable for mapping the genetic basis of diseases and traits in the dog.

Our understanding of how the human genome functions in health and disease will benefit from comparison of its structure with the genomes of other species (*1*, *2*). The domestic dog is a particularly good example, where an unusual population structure offers unique opportunities for understanding the genetic basis of morphology, behaviors, and disease susceptibility (*3*, *4*). The physical and behavioral characteristics of ~300 dog "breeds" are maintained by restricting gene flow between breeds. Many modern breeds are derived from few founders and have been inbred for desired characteristics. This has led to a species with enormous phenotypic diversity, but with significant homogenization of the gene pool within breeds. Many of the ~360 known genetic disorders in dogs resemble human conditions, and their causes may be more tractable in large dog pedigrees than in small, outbred human families (*4*, *5*). The combination of genetic homogeneity and phenotypic diversity also provides an opportunity to understand the genetic basis of many complex developmental processes in mammals (*6*).

Because of the costs of sequencing mammalian genomes to completion, these projects have been restricted to a few species that are considered to be of greatest value to biomedical research. The decision as to whether future projects should aim for complete sequence coverage of a few more genomes, or whether the existing "reference genomes" can be exploited to characterize a wider variety of genomes that are sequenced to a lower level of coverage, must be made. Here,

[1]The Institute for Genomic Research, Rockville, MD 20850, USA. [2]The Center for Advancement of Genomics, Rockville, MD 20850, USA.

*These authors contributed equally to this work.

EXHIBIT B

Combining computational and experimental screening for rapid optimization of protein properties

Robert J. Hayes*, Jörg Bentzien*, Marie L. Ary, Marian Y. Hwang, Jonathan M. Jacinto, Jost Vielmetter, Anirban Kundu, and Bassil I. Dahiyat†

Xencor, 111 West Lemon Avenue, Monrovia, CA 91016

Communicated by Pamela J. Bjorkman, California Institute of Technology, Pasadena, CA, October 16, 2002 (received for review August 14, 2002)

We present a combined computational and experimental method for the rapid optimization of proteins. Using β-lactamase as a test case, we redesigned the active site region using our Protein Design Automation technology as a computational screen to search the entire sequence space. By eliminating sequences incompatible with the protein fold, Protein Design Automation rapidly reduced the number of sequences to a size amenable to experimental screening, resulting in a library of ≈200,000 mutants. These were then constructed and experimentally screened to select for variants with improved resistance to the antibiotic cefotaxime. In a single round, we obtained variants exhibiting a 1,280-fold increase in resistance. To our knowledge, all of the mutations were novel, i.e., they have not been identified as beneficial by random mutagenesis or DNA shuffling or seen in any of the naturally occurring TEM β-lactamases, the most prevalent type of Gram-negative β-lactamases. This combined approach allows for the rapid improvement of any property that can be screened experimentally and provides a powerful broadly applicable tool for protein engineering.

computational protein design | protein engineering | mutagenesis | directed evolution | β-lactamase

The increased use of enzymes and other proteins in the chemical, agricultural, and pharmaceutical industries has generated considerable interest in the design of proteins with new and improved properties. Two different but complementary technologies have been applied to this goal: (*i*) rational design, which relies on structural and mechanistic knowledge and human expertise; and (*ii*) directed evolution methods such as error-prone PCR, phage display, and DNA shuffling, which use random mutagenesis or recombination to create diversity and then experimentally screen the libraries generated for desired properties (1). Directed evolution has been successfully used on a wide range of proteins (2–7). However, this approach is limited by the number of sequences that can be screened experimentally (about 10^{14} for library panning and 10^7 for high-throughput screening). Rational design has also been applied with some success (8–10), but it was not until computational methods were developed that it could be used comprehensively.

Computational techniques use protein design algorithms to perform *in silico* screening of protein sequences (11–17). By taking advantage of the speed of computers, these methods allow a vast number of sequences to be screened ($\approx 10^{80}$). The ability to search such large sequence spaces drastically increases the possibility of finding novel proteins with improved properties. Computational techniques have also been developed to enhance the efficiency of directed evolution methods (18, 19).

One computational design tool that has proven effective is Protein Design Automation (PDA) (13). PDA begins with the three-dimensional structural model of the protein to be designed and predicts the optimal sequence that will adopt this fold, allowing all or a specified set of residues to change. The fitness of sequences is scored by using physical potential functions that model the energetic interactions of protein atoms (20); stable low-energy sequences are given the best scores. By using extremely efficient search algorithms, up to 10^{80} sequences can be accurately screened within hours (21–23). Multiple simultaneous mutations can be made, and novel sequences that are very different from wild type can be discovered. PDA has shown tremendous success in designing proteins with improved stability and conformational specificity (13, 14, 24–28) and has even been used to engineer a catalytic site into a previously nonreactive protein (29).

In these studies, only a few optimal sequences calculated by PDA were made and tested experimentally. The utility of PDA can be extended significantly, however, if it is used to generate a library of sequences, all of which are predicted to be stable and fold into a predetermined structure. Unlike random libraries, where most of the mutations are deleterious, the mutant sequences in the PDA library are computationally screened to eliminate destabilizing mutations and sequences inconsistent with the proper fold. The selected sequences are then experimentally screened for desired properties such as improved catalytic activity, substrate specificity, or receptor binding. Therefore, PDA is a computational prescreen to decrease the sequence space many orders of magnitude, while maintaining broad diversity, to a number easily amenable to experimental screening. By coupling PDA with experimental screening, we combine the advantages of computational design with those of directed evolution: namely, access to a vast sequence space and the ability to improve any protein property that can be captured by a screen.

In this paper, we demonstrate the feasibility of this approach by using it to increase the resistance of bacteria toward the antibiotic cefotaxime by optimizing TEM-1 β-lactamase, the most prevalent plasma-encoded β-lactamase in Gram-negative bacteria.

Methods

Structure Preparation. The crystal structure of TEM-1 β-lactamase (Protein Data Bank no. 1BTL) (30) was used as the starting point for modeling. All water molecules and the sulfate group were removed; the side chains of residues N132, N154, N170, H122, and H289 were flipped to form a better hydrogen bond network; and the disulfide bond between C77 and C123 was formed manually. The program BIOGRAF (Molecular Simulations, San Diego) was used to generate explicit hydrogens, and 50 steps of conjugate gradient minimization were performed by using the Dreiding II force field (31) without the electrostatics term. The minimization is done to make the structure compatible with our force-field parameters and results in very slight changes to the coordinates.

Construction of Mutant Library. To facilitate introduction of the mutations into the *TEM-1* gene, a pCR-Blunt (Invitrogen) vector containing the *TEM-1* gene was digested with *Xba*I and *Hin*dIII,

Abbreviations: PDA, Protein Design Automation; MIC, minimum inhibitory concentration; GMEC, global minimum energy conformation.

*R.J.H. and J.B. contributed equally to this work.

†To whom correspondence should be addressed. E-mail: baz@xencor.com.

www.pnas.org/cgi/doi/10.1073/pnas.212627499

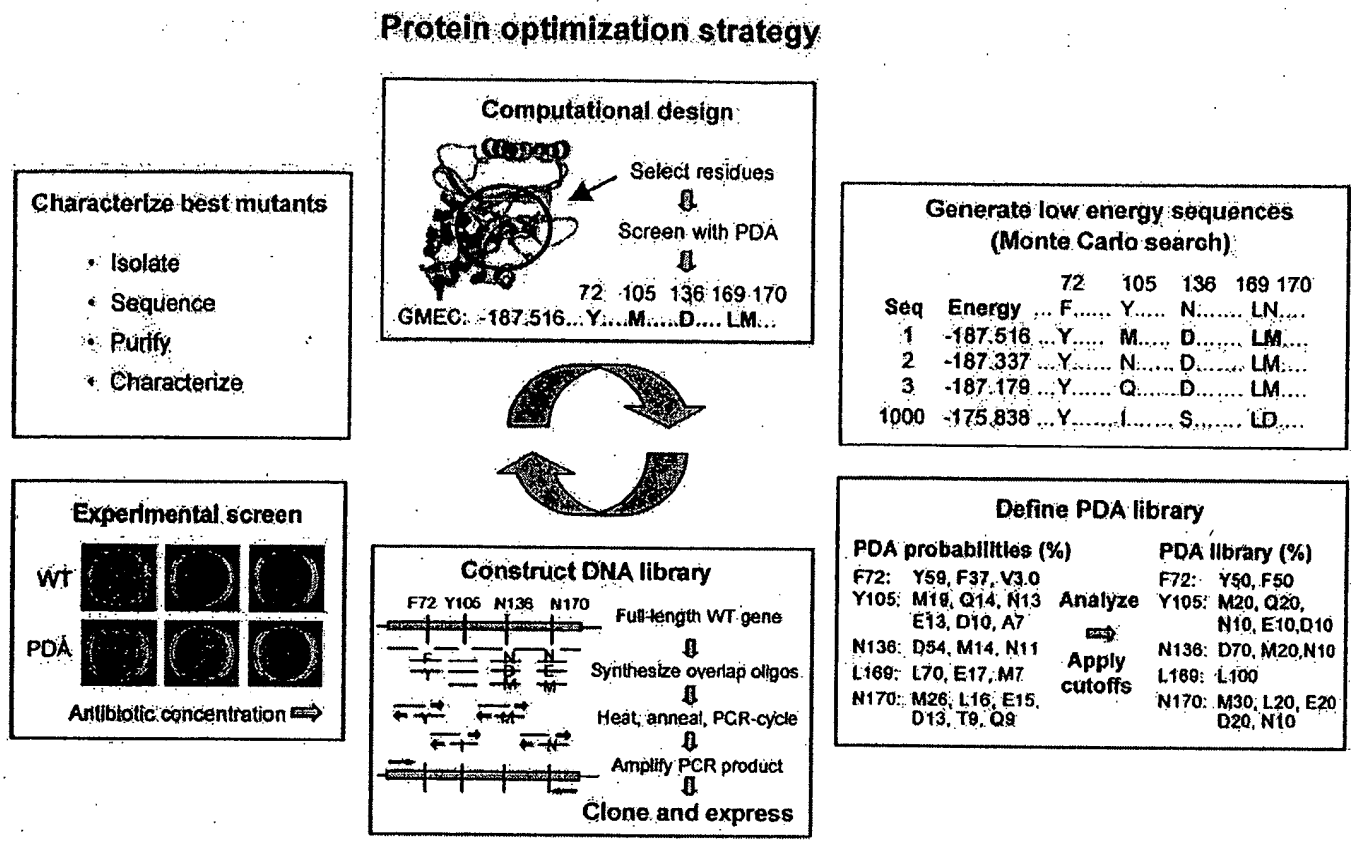


Fig. 1. Protein optimization strategy. A set of residues or region of the protein structure is selected to be designed. PDA is used to computationally screen the entire designed sequence space and determine the GMEC for the fold. Starting from the GMEC, the Monte Carlo search is then used to explore the sequence space and generate a list of near-optimal sequences. An amino acid probability table is obtained from the list and cutoffs are applied to define a PDA library of mutant sequences for experimental screening. The PDA library is translated into a DNA library, which is cloned and expressed. An experimental screen is used to select clones with improved properties; these are then characterized. Results can be fed into additional cycles.

treated with T4 DNA polymerase, and religated. Site-directed mutagenesis was performed by using QuikChange as described by the manufacturer (Stratagene) to remove the existing *Xba*I and *Hin*dIII sites. New *Xba*I and *Hin*dIII sites were then introduced by site-directed mutagenesis at nucleotides 163 and 841, respectively, of the *TEM-1* gene in the vector. A polyhistidine (6×His) sequence was then added to the 3′-end of the *TEM-1* ORF to facilitate immunodetection of the proteins, thereby creating the vector pXR293. The his-tag was found to have no effect on β-lactamase activity. *Escherichia coli* TOP10 cells transformed with pXR293 were confirmed to grow on media containing 100 μg/ml ampicillin and 50 μg/ml kanamycin. The β-lactamase protein expressed from this construct is termed TEM-1 in this report.

The mutated β-lactamase genes were constructed essentially as described by Prodromou and Pearl (32) and Chalmers and Curnow (33). Oligonucleotides corresponding to the gene were synthesized as 40–50 mers with ≈15-nt overlaps. At each mutational position, multiple oligonucleotides were included in the reaction, and the genes were synthesized by using recursive PCR. They were then digested with *Xba*I and *Hin*dIII and subcloned into pXR293. This vector was then transformed into *E. coli* TOP10 cells (Invitrogen) for expression.

Selection of Cefotaxime-Resistant Mutants. *E. coli* cells expressing the mutant library of *TEM-1* genes were grown on plates containing increasing concentrations of cefotaxime, and the minimum inhibitory concentration (MIC) for survival was determined. The cefotaxime concentrations used were: 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, and 128 μg/ml. Assays were conducted at 25 and 30°C to ensure soluble expression of the designed proteins. Cells were plated at low density (<1,000 per plate) to ensure that the observed resistance was not due to confluence. Clones demonstrating the highest resistance were picked, and the β-lactamase protein was identified with immunoblot analysis (34) by using pooled 5- and 6-his polyclonal antibodies (Qiagen, Valencia, CA). The *TEM-1* gene of the most resistant variants was sequenced to identify the mutations. New genes containing the mutations identified for PDA-1, -2, and -3 were constructed, and MICs were determined to confirm the initial screening results.

Results

Protein Optimization Strategy: Combining PDA with Experimental Screening. The overall strategy for protein optimization is shown in Fig. 1. PDA is used to computationally design a protein and define a library of mutant sequences at specific positions. PDA's optimization algorithms are then run to screen all possible sequences for the global optimal sequence and conformation for the target fold, the one with the lowest energy as determined by the scoring function. This conformation is termed the global minimum energy conformation (GMEC). Starting from this optimal structure, a search algorithm such as Monte Carlo (35, 36) simulated annealing is used to explore sequence space and generate a list of other near-optimal sequences. The Monte Carlo list is rank-ordered by energy score and may contain as many sequences as desired (e.g., the best 1,000). An amino acid probability table is then generated from the list by counting

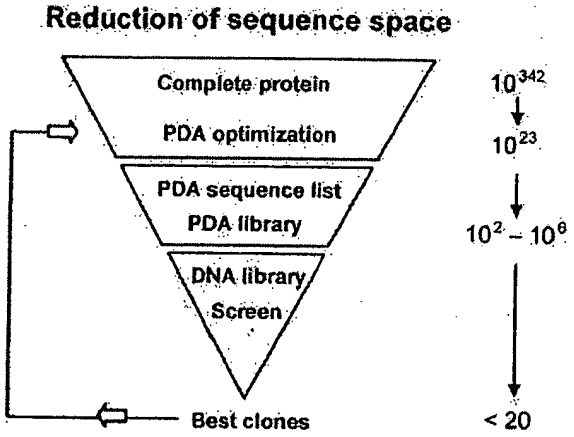


Fig. 2. Reduction of sequence space for PDA design of TEM-1 β-lactamase. Computational screening with PDA and judicious application of cutoffs reduced the sequence space 18 orders of magnitude for the 19 residues explicitly considered and more than 300 orders of magnitude for the entire protein. This conformational screen specified a library for experimental screening of ~200,000 mutant sequences, enriched for structural integrity.

amino acid occurrences at each of the designed positions. Different cutoffs or weighting functions can be applied to define a library of a desired size, appropriate for experimental screening. Structure or sequence alignment information, experimental data, and diversity considerations may also be taken into account in defining the mutant library.

Recursive PCR with overlapping oligonucleotides is then used to synthesize the genes containing all of the mutant sequences in the PDA-defined library. The genes are pooled and cloned, and the mutant proteins are expressed in an appropriate host such as *E. coli*. The mutant proteins are screened experimentally for desired properties, and the best mutants are isolated and characterized. These results can be used as feedback for additional rounds of computational design, library generation, and screening.

Reduction of Sequence Space. The use of PDA as a computational screen allows us to access a vast sequence space and, by eliminating sequences predicted to be destabilizing or inconsistent with the proper fold, reduce it to a size amenable to experimental screening. The reduction of sequence space obtained for TEM-1 β-lactamase, our test case, is shown in Fig. 2. If we were to consider the entire β-lactamase protein (263 residues) and allow all 20 amino acids at each position, we would need to screen 20^{263} or $\approx 1.4 \times 10^{342}$ sequences. By focusing the design to a particular region (19 residues near the active site) and using a slightly restricted set of amino acids (19), we reduced this to 7×10^{23} sequences, a number that can easily be screened computationally, but not experimentally. We then chose cutoffs for the Monte Carlo list and the probability table that would define a library within the limits of experimental screening. In this case, we specified a library of ≈200,000 mutant sequences, a reduction of 18 orders of magnitude for the residues explicitly considered and an overall reduction of more than 300 orders of magnitude for the entire protein.

Computational Design of β-Lactamase. The hydrolysis of β-lactam antibiotics, catalyzed by β-lactamase, is a common mechanism by which bacteria become resistant to antibiotics (37). The most prevalent plasmid-encoded β-lactamase in Gram-negative bacteria is the class A TEM-1 β-lactamase (38). This enzyme hydrolyzes ampicillin efficiently but is inefficient at hydrolyzing the cephalosporin cefotaxime. Our goal was to use PDA to design β-lactamase variants that confer increased resistance toward cefotaxime.

Optimizing the area around the active site is likely to have a significant effect on enzyme activity and substrate specificity (37, 39). Although more distant mutations can also be effective (40), the rationale for how to select such positions is less obvious. Designing residues around the active site also serves as a stringent test of the ability of PDA to predict nondisruptive mutations. We therefore focused our design on residues within 5 Å of the active site residues S70, K73, S130, E166, and K234. These criteria resulted in 19 positions that were allowed to change: M69, T71, F72, V74, V103, Y105, A126, I127, N132, A135, N136, L169, N170, M211, D214, K234, S235, G236, and I247. All 20 amino acids, except cysteine and proline, were considered at these positions. The catalytic residues (S70, K73, S130, and E166) were not allowed to change their amino acid identities; however, their conformations could vary. An expanded version of the backbone-dependent rotamer library of Dunbrack and Karplus (41) was used in all of the calculations, and the DEE algorithm was used to find the GMEC. The computational details, residue classification, and potential functions used are described in previous work (13, 14, 20, 42).

Definition of Mutant Library. Optimization with PDA predicted an optimal sequence with nine mutations. Starting from this GMEC, we applied Monte Carlo simulated annealing to produce a rank-ordered list of the 1,000 lowest energy sequences. A probability table was generated from this list by counting the amino acid occurrences at each of the 19 designed positions (Table 1). A 10% cutoff was then applied to the probability table to define a library of mutant sequences for experimental screening; that is, for a given position, an amino acid identity was included in the library if it had a 10% or greater probability of occurrence. To ensure that the library spanned the complete sequence space from the wild-type enzyme to the most distantly related PDA mutant, we always included the wild-type identity at all designed positions, even if it did not appear in the Monte Carlo list. With a 10% cutoff, this gave us a library of 172,800 unique sequences; a 20% cutoff would have resulted in a much smaller library of 4,806.

Construction of Genes for Mutant Library. Recursive PCR with overlapping oligonucleotides was used to synthesize the *TEM-1* β-lactamase genes containing all 172,800 mutant sequences in the PDA library. Synthetic oligonucleotides containing the designed mutations were pooled to create desired diversity at each site. Two separate reactions were performed: one that contained only a proofreading DNA polymerase (*Pfu* DNA polymerase), termed the nonerror prone reaction, and one that contained both *Pfu* DNA polymerase and *Taq*DNA polymerase, termed the error-prone reaction. The mutated genes were cloned and transformed into *E. coli*.

Validation of Mutant Library. Sixty individual clones from the nonerror-prone library were sequenced by standard techniques. The plasmids contained intact ORFs with the desired mutations. No additional mutations were detected. With a sample size of 60, we were able to find all of the specified mutations at each designed position. It is impossible to find all combinations of the mutations within this small sample (the library contained 172,800 unique sequences), but none of the clones were identical and we were unable to detect a statistically significant bias toward any particular mutation at any position. This result indicates that we have developed an efficient method for converting a PDA-defined library into an experimental library containing all of the mutated genes required to encode the desired mutant sequences.

Experimental Screen for β-Lactamase Activity. Experimental libraries of ≈500,000 individual *E. coli* colonies expressing the mu-

Table 1. PDA probability table for designed positions of TEM-1 β-lactamase

Position	WT	Amino acid probabilities predicted by PDA, %
69	M	D: 81.2 A: 7.6 G: 10.7 S: 1.5
71	T	T: 100.0
72	F	F: 87.5 Y: 12.5
74	V	V: 85.2 L: 8.8 D: 3.0 I: 1.7
103	V	Q: 82.1 K: 11.1 F: 2.0
105	Y	Y: 83.3 N: 10.9 I: 12.2 S: 10.4 E: 8.3 D: 2.4
126	A	A: 95.6 S: 4.4
127	I	I: 74.2 L: 14.6
132	N	N: 97.1 L: 2.9
135	A	A: 93.2 G: 3.9 S: 2.8
136	N	N: 84.6 D: 9.7 V: 5.4
169	L	A: 81.7 E: 13.6 D: 2.6 S: 2.1
170	N	N: 95.1 M: 4.0 F: 1.0
211	M	M: 96.6
214	D	D: 100.0
234	K	M: 20.6 L: 25.5 I: 24.1 E: 14.8 Q: 3.0 V: 2.6
235	S	D: 70.7 A: 21.9
236	G	G: 92.5 S: 16.1 A: 2.4
247	I	I: 99.3

Amino acids included when applying different probability cutoffs are indicated as follows: 20% cutoff in dark gray, 10% cutoff in light gray, and 1% cutoff in white background. In defining the PDA library, the 10% cutoff was used and the wild-type amino acid identity was added if it did not appear in the Monte Carlo list. This specified a library of 172,800 unique sequences.

tated β-lactamase genes were pooled and plated onto increasing concentrations of cefotaxime in a single round of selection, and the MIC for survival was determined. This number of colonies is about three times the size of the PDA-defined library and was used to ensure that the pooled DNA library contained at least one copy of each mutant sequence (within a 95% level of confidence) (43). Clones from the nonerror-prone library had a MIC of 64 μg/ml, which is a 640-fold increase in resistance compared with the wild-type value of 0.1 μg/ml. Clones from the error-prone library had a MIC of 128 μg/ml, a 1,280-fold increase in resistance (see PDA-1 and -2, Table 2). Because our approach allows us to assay the complete PDA library diversity

Table 2. Antibiotic resistance of TEM-1 β-lactamase variants

Variant	Mutations	No. of mutations	No. of novel mutations	Cefotaxime		Ampicillin	
				MIC, μg/ml	Fold increase*	MIC, μg/ml	Fold decrease*
TEM-1 (WT)	—	—	—	0.1	—	4,096†	—
PDA-1	**M69D, V103Q, Y105N N132M, L169A, N170L S235D, G236S**	8	8	64	640	100‡	40
PDA-2	**V103Q, Y105N, I127L L169A, S235Y, G236S**	6	6	128	1,280	100‡	40
PDA-3	**M69D, V103Q, N132M L169A, N170L, S235D**	6	6	64‡	640	ND	ND
TEM-15	**E104K, G238S**	2	—	16	160	4,096§	1
ST-1	**E104K, G238S, M182T A18V**	4	—	256	2,560	ND	ND

Resistance measured at 25°C unless specified otherwise. ND, MIC not determined.
*Fold increase/decrease in resistance is relative to wild type (TEM-1).
†Value reported by Cantu and Palzkill (44).
‡MIC assay done at 30°C.
§Value reported by Shannon *et al.* (49). Boldface indicates novel mutations (not reported to significantly improve cefotaxime resistance in any TEM-β-lactamase; refs. 2, 45, and 46; Jacoby, G. and Bush, K., www.lahey.org/studies/temtable.htm).

in a single round, we were able to use very stringent selection conditions and directly obtain highly resistant variants. The identification of incrementally improved sequences was not necessary.

Substrate Specificity. We also measured resistance to ampicillin and found no growth at 100 μg/ml, significantly less than the MIC of 4,096 μg/ml reported for the wild type (43). This result suggests that our screens identified clones whose resistance to cefotaxime had dramatically improved, whereas their resistance to ampicillin was reduced at least 40-fold. The relative substrate specificity toward cefotaxime vs. ampicillin was thus enhanced 25,000- to 50,000-fold.

PDA Mutants Are Novel. The most active mutant from each library was isolated and sequenced. PDA-1 had eight mutations (M69D, V103Q, Y105N, N132M, L169A, N170L, S235D, and G236S), all designated in the PDA library (see Table 1). PDA-2 had five PDA-designed mutations (V103Q, Y105N, I127L, L169A, and G236S) and one random mutation (S235Y). The S235Y mutation was not predicted by PDA due to steric clashes. Protein backbone motion, which is required to relieve the clash, is not considered in the computation. None of the mutations in PDA-1 or -2 have been identified by full gene random mutagenesis or DNA shuffling studies (2, 39, 45, 46) or have been observed in the 105 naturally occurring TEM β-lactamases (G. Jacoby and K. Bush, www.lahey.org/studies/temtable.htm). Orencia *et al.* (47) discussed the emergence of antibiotic resistances in β-lactamases and showed that there is an overlap between the mutations discovered by directed evolution and those occurring in natural evolution. PDA, however, accesses the entire designed sequence space including all possible combinations of mutations and therefore can produce multiple simultaneous mutations. PDA is therefore more likely to identify novel mutants with desired properties. The lone random mutation in PDA-2 (S235Y) was in the active site region, suggesting that the novel context of the PDA-designed mutants allowed this previously unobserved, but beneficial, mutation to emerge.

Two of the mutations in PDA-1 (V105N and G236S) were reverted to wild type to create a backcross mutant (PDA-3). This PDA-3 sequence is present in the library defined with a 20% cutoff but is absent if a 10% cutoff is used (see Table 1). PDA-3 exhibited the same cefotaxime resistance as PDA-1 (Table 2), indicating that a smaller PDA-library (4,806 vs. 172,800 sequences) can also generate mutants with significantly improved activity. Additional backcrosses were done to examine the role of the other six mutations in PDA-1. No single mutation was primarily responsible for the improved resistance, and no simple additivity was apparent, suggesting that the mutations are coupled. This conclusion is supported by extensive replacement mutagenesis studies of three-residue segments around the active site (39). They found a mutant (E168G, L169A, and N170G) that included one of our mutations (L169A), but it showed only a marginal (2-fold) improvement in cefotaxime resistance. Although they also tested most of our other mutations, no increased resistance was found for any of these. This lack of improved resistance indicates that the broader context of many simultaneous mutations provided by our approach was required to find our highly active sequences.

Comparison with Other Mutants. To compare the activity of our PDA-designed mutants with those obtained in other studies, we introduced some previously reported mutations into our wild-type gene, including E104K/G238S (comparable to TEM-15) (2, 46) and A18V/E104K/M182T/G238S (comparable to ST-1) (2, 46). TEM-15 is a naturally occurring β-lactamase that is active against cefotaxime, and ST-1 is a highly active TEM-1 variant discovered from three rounds of DNA shuffling. We tested the ability of these mutants to confer resistance to cefotaxime. Wild type had a MIC of 0.1 μg/ml, comparable to the values reported by others; TEM-15 and ST-1 had MICs of 16 and 256 μg/ml, respectively, also in line with previously published work (Table 2) (2, 44, 46, 48–50).

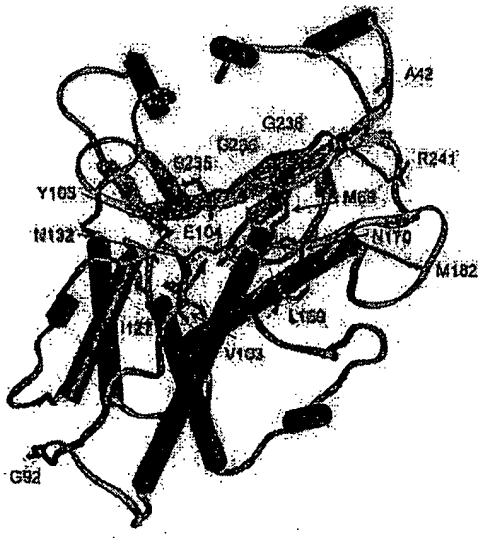


Fig. 3. Location of mutations in PDA-1 and -2 (green) vs. those obtained by DNA shuffling (2) and random hypermutagenesis (46) (magenta). The wild-type TEM-1 β-lactamase structure is illustrated, and the side chains of the mutated positions are shown. The catalytic serine (S70) is depicted in blue. The average distance between the C_α atoms and the catalytic nucleophile (O,S70) in our PDA-1 and -2 mutations was 8.0 and 8.6 Å, respectively, vs. 16.0 Å for the mutations in ST-2 and -3 (Stemmer's best mutants) (2) and 12.1 Å for 3D.5 (Zaccolo and Gherardi's best mutant) (46). This difference in distances illustrates that the mutations found by PDA are near the designed active site area, whereas those found by DNA shuffling and random hypermutagenesis are farther away.

Location of Mutations. The mutations in all our variants are located in or near the active site, because our computational design restricted changes to this region. Directed evolution methods, however, tend to produce mutations spread over the entire protein structure. For example, almost all of the mutations in the best mutants obtained by DNA shuffling (2) and random hypermutagenesis (46) are located far from the active site (Fig. 3). It is possible that these techniques seldom produce mutations close to the active site, because they rely on incremental changes; a single change in the first round of screening must be beneficial to be passed to the second round. However, point mutations in the active site area are usually disruptive. Our approach, on the other hand, allows multiple simultaneous mutations in a single round, which can have compensating or even synergistic effects.

Sequence Space Coverage. Most of the mutations observed in our PDA variants require a minimum of two nucleotide changes, and one, M69D, can be made only by a triple nucleotide change (Table 3). Double- or triple-nucleotide changes within a single codon are very difficult to achieve by using random mutagenesis techniques such as error-prone PCR or single-gene DNA shuffling. This limitation is demonstrated by the fact that each of the mutations found in the directed evolution studies (2, 45, 46) as well as those observed in the 105 naturally occurring TEM variants (G. Jacoby and K. Bush, www.lahey.org/studies/temtable.htm) could be obtained by a single nucleotide change. If one considers all of the substitutions that are possible for each of the 20 amino acids, on average only seven can be achieved by a single nucleotide change. The sequence space coverage is further reduced by codon preferences, biases for transitions over transversions, and A $\leftrightarrow$ T over G $\leftrightarrow$ C mutations. These restrictions severely limit the sequence

Table 3. Minimal number of nucleotide changes required for amino acid mutations in TEM-1 β-lactamase variants

Variant	Position																	
	18	42	69	92	103	104	105	127	132	169	170	182	235	236	238	240	241	254
TEM-1 (WT)	A	A	M	G	V	E	Y	I	N	L	N	M	S	G	G	E	R	D
TEM-15						K,1									S,1			
ST-1*	V,1					K,1						T,1			S,1			
ST-2*†, ST-3*		G,1		S,1		K,1						T,1			S,1		H,1	
ST-4*						K,1						T,1			S,1			
3D.5†‡						K,1						T,1			S,1			
3A.6†‡															S,1	K,1	H,1	G,1
PDA-1			D,3		Q,2		N,1		M,2	A,2	L,2		D,2	S,1				
PDA-2					Q,2		N,1	L,1		A,2			Y,1	S,1				
PDA-3			D,3		Q,2				M,2	A,2	L,2		D,2					

*Stemmer (2).
†Has additional silent mutations.
‡Zaccolo and Gherardi (46). Mutations requiring two or more nucleotide changes are shown in bold.

space accessible to these methods and suggest why our approach, which does not suffer from these limitations, is more likely to produce novel functional sequences.

Discussion

The purpose of this study is to show that a combined approach, using PDA as a computational screen to rationally reduce the sequence space before experimental screening, can rapidly lead to novel protein variants with improved properties. We used PDA to identify sequences compatible with the protein fold and then experimentally screened the resulting sequence library to obtain variants with novel properties. As a test case, we redesigned the active site of β-lactamase and then selected for variants with improved resistance to cefotaxime. In a single round, we obtained variants that exhibit a 1,280-fold increase in resistance and, to our knowledge, are novel.

Our approach has some key distinctions from purely experimental techniques. By using an efficient computational screen, we are able to access an extremely large region of sequence space and rapidly reduce it to a number amenable to experimental screening. Experimental libraries will always be restricted to sampling a miniscule portion of sequence space due to limitations on the sheer mass of proteins that can be physically made and tested. Computational screening, on the other hand, is scalable and can comprehensively screen enormous sequence spaces. The sequence space searched by random mutagenesis techniques is also severely limited by the fact that amino acid mutations requiring more than single nucleotide changes are extremely unlikely. Random techniques usually produce only incremental changes per round and require multiple rounds, whereas our approach creates multiple simultaneous mutations in a single round. All of these features result in the rapid discovery of novel mutants that are different from any of those observed previously.

Another difference with the PDA method is that it offers full control over the location and type of mutations, allowing incorporation of structural and experimental data and enhancing our understanding of structure–activity relationships. It results in focused designs of smaller functionally enriched mutant libraries, allowing complex and expensive screens such as mammalian cell-based assays for improved protein therapeutics, while still accessing broad sequence diversity. Our combined computational and experimental approach allows for the rapid improvement of any protein property that is amenable to experimental screening and has broad applications in the chemical, agricultural, and pharmaceutical industries.

1. Arnold, F. H. (2001) *Nature* **409**, 253–257.
2. Stemmer, W. P. (1994) *Nature* **370**, 389–391.
3. Arnold, F. H. & Moore, J. C. (1997) *Adv. Biochem. Eng. Biotechnol.* **58**, 1–14.
4. Crameri, A., Raillard, S. A., Bermudez, E. & Stemmer, W. P. (1998) *Nature* **391**, 288–291.
5. Zhao, H., Giver, L., Shao, Z., Affholter, J. A. & Arnold, F. H. (1998) *Nat. Biotechnol.* **16**, 258–261.
6. Minshull, J. & Stemmer, W. P. (1999) *Curr. Opin. Chem. Biol.* **3**, 284–290.
7. Pedelacq, J. D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C. Y., Rho, B. S., Park, M. S. Terwilliger, T. C. & Waldo, G. S. (2002) *Nat. Biotechnol.* **29**, 927–932.
8. Cunningham, B. C. & Wells, J. A. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 3407–3411.
9. Whittle, E. & Shanklin, J. (2001) *J. Biol. Chem.* **276**, 21500–21505.
10. Bornscheuer, U. T. & Pohl, M. (2001) *Curr. Opin. Chem. Biol.* **5**, 137–143.
11. Hellinga, H. W. & Richards, F. M. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 5803–5807.
12. Desjarlais, J. R. & Handel, T. M. (1995) *Protein Sci.* **4**, 2006–2018.
13. Dahiyat, B. I. & Mayo, S. L. (1996) *Protein Sci.* **5**, 895–903.
14. Dahiyat, B. I. & Mayo, S. L. (1997) *Science* **278**, 82–87.
15. Street, A. G. & Mayo, S. L. (1999) *Struct. Folding Des.* **7**, R105–R109.
16. Kraemer-Pecore, C. M., Wollacott, A. M. & Desjarlais, J. R. (2001) *Curr. Opin. Chem. Biol.* **5**, 690–695.
17. Pokala, N. & Handel, T. M. (2001) *J. Struct. Biol.* **134**, 269–281.
18. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. G. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 3778–3783.
19. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002) *Nat. Struct. Biol.* **3**, 1–6.
20. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999) *Curr. Opin. Struct. Biol.* **9**, 509–513.
21. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992) *Nature* **356**, 539–542.
22. Gordon, D. B. & Mayo, S. L. (1998) *J. Comp. Chem.* **10**, 1505–1514.
23. Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000) *J. Comp. Chem.* **21**, 999–1009.
24. Malakauskas, S. M. & Mayo, S. L. (1998) *Nat. Struct. Biol.* **5**, 470–475.
25. Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L. & Springer, T. A. (2000) *Nat. Struct. Biol.* **7**, 674–678.
26. Marshall, S. A. & Mayo, S. L. (2001) *J. Mol. Biol.* **305**, 619–631.
27. Luo, P., Hayes, R. J., Chan, C., Stark, D. M., Hwang, M. Y., Jacinto, J. M., Juvvadi, P., Chung, H. S., Kundu, A., Ary, M. L. & Dahiyat, B. I. (2002) *Protein Sci.* **11**, 1218–1226.
28. Filikov, A. V., Hayes, R. J., Luo, P., Stark, D. M., Chan, C., Kundu, A. & Dahiyat, B. I. (2002) *Protein Sci.* **11**, 1452–1461.
29. Bolon, D. N. & Mayo, S. L. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 14274–14279.
30. Jelsch, C., Mourey, L., Masson, J. M. & Samama, J. P. (1993) *Proteins* **16**, 364–383.
31. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94**, 8897–8909.
32. Prodromou, C. & Pearl, L. H. (1992) *Protein Eng.* **5**, 827–829.
33. Chalmers, F. M. & Curnow, K. M. (2001) *Biotechniques* **30**, 249–252.
34. Harlow, E. & Lane, D. (1988) *Antibodies: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 471–510.
35. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21**, 1087–1092.
36. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000) *J. Mol. Biol.* **299**, 789–803.
37. Matagne, A., Lamotte-Brasseur, J. & Frere, J. M. (1998) *Biochem. J.* **330**, 581–598.
38. Sutcliffe, J. G. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3737–3741.
39. Palzkill, T. & Botstein, D. (1992) *J. Bacteriol.* **174**, 5237–5243.
40. Spiller, B., Gershenson, A., Arnold, F. H. & Stevens, R. C. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 12305–12310.
41. Dunbrack, R. L., Jr., & Karplus, M. (1993) *J. Mol. Biol.* **230**, 543–574.
42. Street, A. G. & Mayo, S. L. (1998) *Folding Des.* **3**, 253–258.
43. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
44. Cantu, C., III, & Palzkill, T. (1998) *J. Biol. Chem.* **273**, 26603–26609.
45. Vakulenko, S. B., Geryk, B., Kotra, L. P., Mobashery, S. & Lerner, S. A. (1998) *Antimicrob. Agents Chemother.* **42**, 1542–1548.
46. Zaccolo, M. & Gherardi, E. (1999) *J. Mol. Biol.* **285**, 775–783.
47. Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. & Stevens, R. C. (2001) *Nat. Struct. Biol.* **8**, 238–242.
48. Sirot, D., Recule, C., Chaibi, E. B., Bret, L., Croize, J., Chanal-Claris, C., Labia, R. & Sirot, J. (1997) *Antimicrob. Agents Chemother.* **41**, 1322–1325.
49. Shannon, K., Stapleton, P., Xiang, X., Johnson, A., Beattie, H., El Bakri, F., Cookson, B. & French, G. (1998) *J. Clin. Microbiol.* **36**, 3105–3110.
50. Stapleton, P. D., Shannon, K. P. & French, G. L. (1999) *Antimicrob. Agents Chemother.* **43**, 1881–1887.

EXHIBIT C

Development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening

PEIZHI LUO, ROBERT J. HAYES, CHERYL CHAN, DIANE M. STARK, MARIAN Y. HWANG, JONATHAN M. JACINTO, PADMAJA JUVVADI, HELEN S. CHUNG, ANIRBAN KUNDU, MARIE L. ARY, AND BASSIL I. DAHIYAT
Xencor, Inc., Monrovia, California 91016, USA

(RECEIVED November 13, 2001; FINAL REVISION February 11, 2002; ACCEPTED February 13, 2002)

Abstract

Granulocyte-colony stimulating factor (G-CSF) is used worldwide to prevent neutropenia caused by high-dose chemotherapy. It has limited stability, strict formulation and storage requirements, and because of poor oral absorption must be administered by injection (typically daily). Thus, there is significant interest in developing analogs with improved pharmacological properties. We used our ultrahigh throughput computational screening method to improve the physicochemical characteristics of G-CSF. Improving these properties can make a molecule more robust, enhance its shelf life, or make it more amenable to alternate delivery systems and formulations. It can also affect clinically important features such as pharmacokinetics. Residues in the buried core were selected for optimization to minimize changes to the surface, thereby maintaining the active site and limiting the designed protein's potential for antigenicity. Using a structure that was homology modeled from bovine G-CSF, core designs of 25–34 residues were completed, corresponding to 10^{21}–10^{28} sequences screened. The optimal sequence from each design was selected for biophysical characterization and experimental testing; each had 10–14 mutations. The designed proteins showed enhanced thermal stabilities of up to 13°C, displayed five- to 10-fold improvements in shelf life, and were biologically active in cell proliferation assays and in a neutropenic mouse model. Pharmacokinetic studies in monkeys showed that subcutaneous injection of the designed analogs results in greater systemic exposure, probably attributable to improved absorption from the subcutaneous compartment. These results show that our computational method can be used to develop improved pharmaceuticals and illustrate its utility as a powerful protein design tool.

Keywords: Protein design; computational screen; stability; cytokines; granulocyte-colony stimulating factor

Many techniques have been used in the design of new and improved proteins. In vitro directed evolution methods such as phage display, DNA shuffling, and error-prone PCR are widely used. Rational design approaches continue to be applied, and strategies that combine both are now being used. Successful designs include enzymes (Chen and Arnold 1991; Stemmer 1994; Zhao et al. 1998) and other proteins (Crameri et al. 1996), as well as therapeutically useful proteins such as hormones and cytokines (Lowman and Wells 1993; Heikoop et al. 1997; Grossmann et al. 1998; Chang et al. 1999). The experimental techniques involve the generation and screening of libraries of random protein sequences. However, the number of sequences that can be screened experimentally is limited (about 10^{14} for library panning and 10^7 for high throughput screening). Libraries of this size allow for the simultaneous modification of only about 10 residues.

Reprint requests to: Bassil I. Dahiyat, Xencor, Inc., 111 W. Lemon Avenue, Monrovia, California 91016, USA; e-mail: baz@xencor.com; fax: (626) 256-3562.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.4580102.

Computational methods have also been used that perform in silico screening of protein sequences (Hellinga and Richards 1994; Desjarlais and Handel 1995; Dahiyat and Mayo 1996, 1997a; Street and Mayo 1999; Jiang et al. 2000; Kraemer-Pecore et al. 2001; Pokala and Handel 2001). Exploiting the efficiency and speed of computers, these methods can randomly screen a vast number of sequences (up to 10^{80}), allowing for the simultaneous consideration and modification of more than 60 residues. Searching such large sequence spaces drastically improves the possibility of finding novel protein sequences with improved properties.

Investigators have recently developed a computational screening method that finds the optimal sequence for a defined three-dimensional structure, allowing all or part of the sequence to change (Dahiyat and Mayo 1996). This method, termed Protein Design Automation (PDA), scores the fit of sequences to the three-dimensional structure using physical-chemical potential functions that model the energetic interactions of protein atoms, including steric, solvation, and electrostatic interactions. PDA couples these potential functions with a highly efficient search algorithm to accurately screen up to 10^{80} sequences. Because the screening is performed in silico, multiple simultaneous mutations can be made, and novel sequences that are very different from wild type can be discovered. The method has been validated by numerous experimental tests and has resulted in the design of new proteins with improved stability and conformational specificity, and novel activity (Dahiyat and Mayo 1996, 1997a; Malakauskas and Mayo 1998; Strop and Mayo 1999; Shimaoka et al. 2000; Bolon and Mayo 2001; Marshall and Mayo 2001).

PDA also has the advantage of being able to control the location and type of mutations. For example, the design can be limited to the hydrophobic core. Mutations in the core can produce significant improvements in protein stability but do not change binding epitopes on the surface of the molecule. Thus, the molecular surface can be kept identical to the native structure, retaining biological activity and limiting toxicity and antigenicity. This feature is particularly important in the design of therapeutic proteins.

We wanted to take advantage of these features of PDA and explore its utility in the design of improved pharmaceuticals. We therefore used PDA as an ultrahigh throughput screen for improved analogs of a therapeutic protein, granulocyte-colony stimulating factor (G-CSF). G-CSF is a hematopoietic growth factor of 174 residues that induces differentiation and proliferation of granulocyte-committed progenitor cells. It is used clinically to treat cancer patients and alleviate the neutropenia induced by high-dose chemotherapy. G-CSF belongs to the class of long-chain four-helix bundle cytokines that bind asymmetrically to homodimeric complexes of cell-surface receptors to initiate an intracellular signaling cascade. Their structural similarity allows the design strategy chosen for G-CSF to be immediately applicable to the other four-helix bundle cytokines (human growth hormone, erythropoietin, the interleukins, and interferon-α/β—all clinically important compounds) and thus broadens the potential impact of the results.

Although the cytokines are functionally very efficacious, their pharmacological properties are not ideal. For example, G-CSF, like most proteins, is not absorbed orally to any significant extent and must be administered by frequent (daily) injections throughout the course of treatment. It also has limited stability and strict formulation and storage requirements, including the need to be kept refrigerated. Thus, there is significant interest in developing analogs with improved pharmacological properties.

We sought to use PDA to improve the physicochemical characteristics of G-CSF. Improving these properties can make a molecule more robust, enhance its shelf life, or make it more amenable to use in alternate delivery systems and formulations. It can also affect clinically important features such as pharmacokinetics and result in a drug that is safer for human use. Our design strategy was to optimize the core to improve the stability and solution properties of G-CSF while preserving receptor binding and biological activity.

The template structure used for in silico screening was a homology model of human G-CSF in which the human sequence was mapped onto bovine G-CSF. We designed several novel core sequences, cloned and expressed them, characterized their stabilities, tested them for functional activity both in vitro and in vivo, and studied their pharmacokinetics in monkeys. The designed proteins showed enhanced thermal stabilities, displayed five- to 10-fold improvements in shelf life, and were biologically active both in cell proliferation assays and in a neutropenic mouse model. Subcutaneous injection of the most stable variant in monkeys also resulted in greater systemic exposure, probably attributable to improved absorption from the subcutaneous compartment. These results indicate that PDA has great potential as a powerful in silico tool in the design of improved pharmaceutical proteins.

Results and Discussion

Homology modeling

The crystal structure of bovine G-CSF (PDB record 1bgc) (Lovejoy et al. 1993) was used as the starting point for modeling because the crystal structure of human G-CSF (PDB record 1rhg) (Hill et al. 1993) is at a lower resolution and is missing key fragments, including a structurally important disulfide bond between positions 64 and 74. Bovine G-CSF is a good model for human G-CSF because the sequences are the same length and 142 of 174 amino acids are identical (82%). The residues that differ in the bovine sequence were replaced with the human residues for those

positions, and the conformations of the replaced side chains were optimized using PDA. Most of the replaced residues were solvent exposed, thereby introducing little strain into the structure and allowing typical PDA parameters to be used for conformation optimization. One substitution, however, was at a buried site, G167V, and clashed sterically with a nearby disulfide bond. To accommodate the larger Val, the side-chain conformation at this position was optimized using a less restrictive van der Waals scale factor (0.6 instead of 0.9). The entire structure was then briefly minimized to relax the strain. The final structure that served as the template for all the designs is shown in Figure 1.

Core designs

Unlike many experimental sequence screening methods, PDA allows control over which residues are allowed to change. Core residues were selected because optimization of these positions can improve stability yet minimize changes to the molecular surface, thus limiting the designed protein's potential for antigenicity. Ala scanning studies of G-CSF indicate one or two binding sites on the protein surface that are probably responsible for granulopoietic activity (Reidhaar-Olson et al. 1996; Young et al. 1997) (Fig. 1). Although recent crystallographic studies of G-CSF complexed to its receptor show only one binding site in a novel 2:2 complex (Horan et al. 1996; Aritomi et al. 1999), both sites were avoided in the core designs to ensure preservation of function.

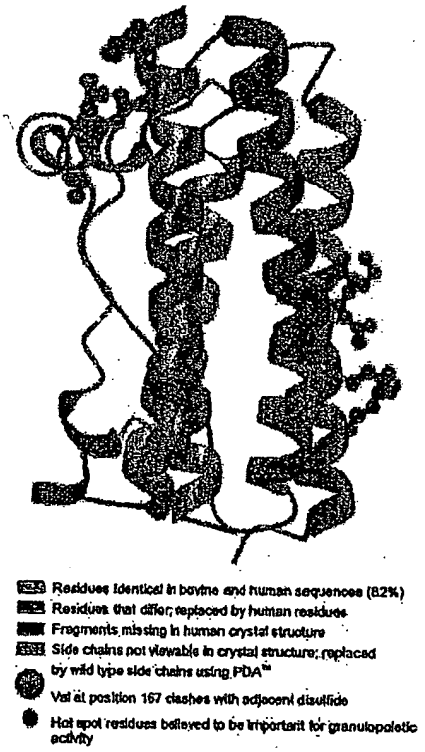


Fig. 1. Template structure of hG-CSF used for Protein Design Automation (PDA) designs. The human sequence was homology modeled onto the bovine crystal structure (PDB record 1bgc). The residues that differ in the bovine sequence or were not present in the bovine crystal structure were replaced with the residues from the human sequence. The conformations of the replaced side chains were optimized using PDA (the larger Val at position 167 was optimized using a less restrictive van der Waals scale factor), and the entire structure was energy minimized for 50 steps.

Two PDA design calculations were run: a deep core design that included residues deeply buried in the interior of the protein and an expanded core design (exp_core) that also included less buried peripheral core residues. The deep core design had 26 core positions that were allowed to vary (shown yellow and gold in Fig. 2), whereas exp_core had 34 (shown yellow and turquoise in Fig. 2). Only hydrophobic amino acids were considered at the variable core positions. These included Ala, Val, Ile, Leu, Phe, Tyr, and Trp. Gly was also allowed for the variable positions that had Gly in the bovine wild-type structure (positions 28, 149, 150, and 167). Met and Pro were not allowed.

Optimal sequences

The optimal sequences selected by PDA are also shown in Figure 2. The optimal sequence from the deep core design had 10 mutations (named core10), and the optimal exp_core sequence had 11 (named exp_core11); thus, 33%–38% of the variable residues changed their identities. Eight of the mutated positions changed to the same amino acid in both designs. Changing the set of design positions can significantly impact the amino acid selected at a given position. For example, in the deep core design, Leu89 retains the same amino-acid identity and conformation as wild type. However, in the exp_core design, when Leu92 is also allowed to vary, both positions (Leu89 and Leu92) mutate to Phe, indicating a coupling between these two core residues. The modeled structure of the sequence selected in the deep core design (core10) is shown in Figure 3.

Native human G-CSF (met hG-CSF) and the optimal sequence from each of the core designs were cloned, expressed in *Escherichia coli*, and purified for experimental studies.

Thermal stability

The far-ultraviolet (UV) circular dichroism (CD) spectra for met hG-CSF and the designed proteins were nearly identical to each other and to published spectra for met hG-CSF (Reidhaar-Olson et al. 1996; Young et al. 1997), indicating highly similar secondary structure and tertiary folds (data

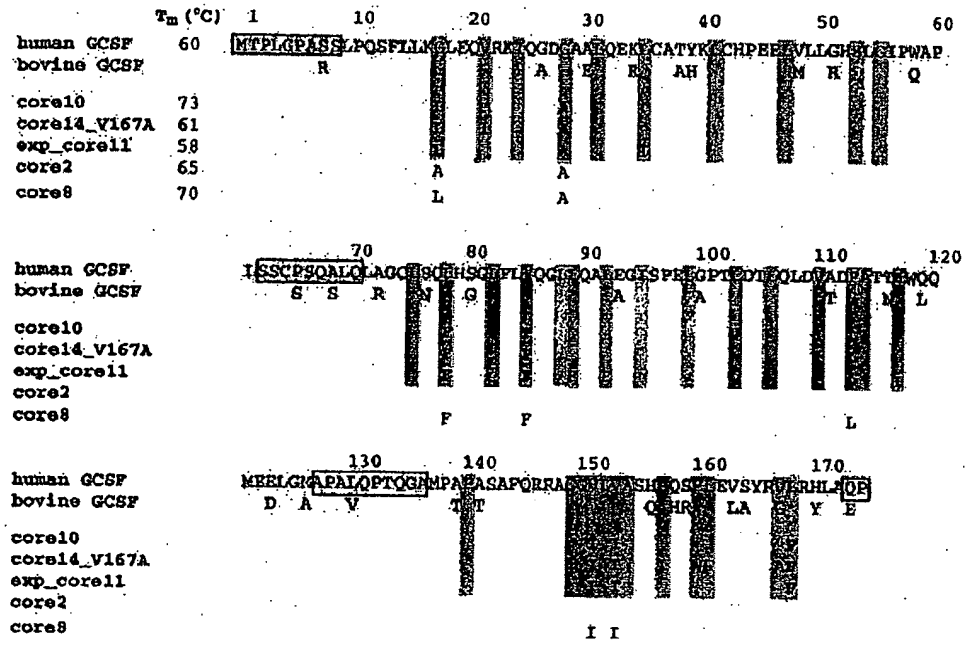


Fig. 2. Sequences of hG-CSF analogs. Native human and bovine sequences are shown at the *top*. The fragments missing in the crystal structure of the human sequence are shown boxed. Variable positions are colored. The deep core design had 26 variable positions, exp_core had 34, and core167V had 25. The optimal sequence from each design is shown. Letters indicate core residues that mutated relative to native hG-CSF; blanks indicate no change. Positions that changed to the same amino acid in all three core designs are indicated in bold. Core2 and core8 sequences were not obtained from PDA calculations but were derived by reverting some of the core10 mutations to wild type. Melting temperatures (T_ms) obtained for the designed proteins are also shown.

not shown). Thermal denaturation was monitored at 222 nm, and the melting temperatures (T_ms) were derived from the derivative curve of the ellipticity at 222 nm versus temperature (Fig. 4). Thermal denaturation of G-CSF and its variants is irreversible; however, T_m can be used to quickly assess the relative stability of different mutants. Stability under storage conditions, which is more relevant clinically, was evaluated with shelf-life studies (see below).

The T_m for met hG-CSF was 60°C, identical to that reported in other studies (Kolvenbach et al. 1997). Core10 showed an increase in stability of 13°C, whereas the T_m of exp_core11 was very similar to wild type (Fig. 2 and Fig. 4). The increased stability seen with core10 may be attributable to improved packing interactions and optimized hydrophobic burial of side chains. Other possibilities include decreased aggregation resulting from elimination of the free cysteine at position 17. The Gly to Ala mutation at position 28 caused a significant improvement in helical propensity that could also be the source of the improved stability.

Identifying critical mutations using derived sequences

To differentiate between these possibilities, two additional sequences derived from the core10 mutant sequence were made and their T_ms measured. One of these (core8) was identical to core10 except that two mutations distant from the others were reverted to wild type (L103V and V110I). These were the two positions that did not mutate in exp_core11. The T_m of core8 was 70°C, similar to core10, indicating that the mutations at 103 and 110 were not responsible for core10's improved stability.

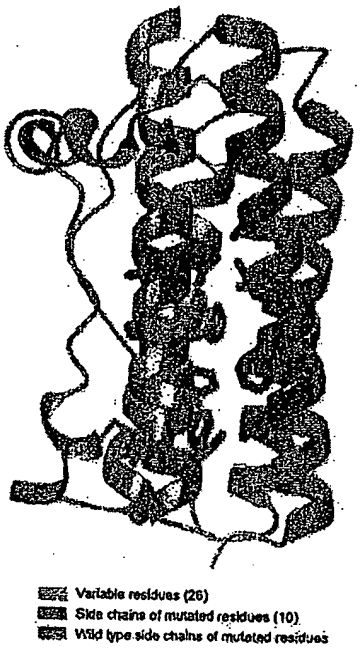


Fig. 3. Modeled structure of hG-CSF analog (core10) obtained from deep core design. Twenty-six core residues were allowed to vary; computational screening with PDA resulted in 10 mutations: C17L, G28A, L78F, Y85F, L103V, V110I, F113L, V151I, V153I, and L168F.

To determine the importance of the other mutations, another sequence was made (core2) that contained only two of the core10 mutations, G28A and C17A; all other residues were identical to wild type (Fig. 2). The T_m of core2 was 5°C higher than wild type, indicating that improvements in helical propensity and the elimination of a free cysteine are important for heightened thermostability. The remainder of the increase in T_m seen for core10 may be attributable to improved packing interactions and increased hydrophobic burial.

Storage stability

Increased shelf life is important for distribution and storage and is a desirable feature for G-CSF and other protein drugs. Because aggregation and chemical degradation are the predominant mechanisms of inactivation of G-CSF (Herman et al. 1996), shelf life was estimated by incubating the proteins at elevated temperature and then using size-exclusion chromatography to observe the disappearance of monomeric protein. Chemical degradation was estimated using reverse phase chromatography (data not shown). Core2 and core10 showed five- and 10-fold improvements in storage stability, respectively, at 50°C (Fig. 5). Rate constants were determined by a first order exponential fit of the fraction monomer remaining/time curves using KaleidaGraph (Synergy Software).

Biological activity

Granulopoietic activity was determined in vitro by quantitating cell proliferation as a function of protein concentration in murine lymphoid cells transfected with the gene for the human G-CSF receptor. The designed proteins were as

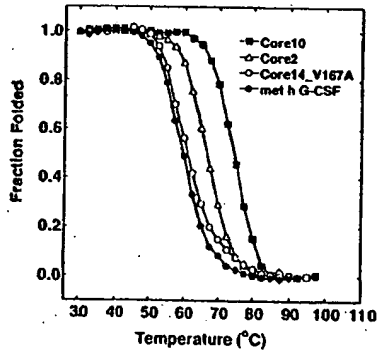


Fig. 4. Thermal stability of hG-CSF analogs. Thermal stability was assessed by monitoring the temperature dependence of the circular dichroism spectral signal at 222 nm. Melting temperatures (T_ms) were derived from the derivative curve of the ellipticity at 222 nm versus temperature. Core10 and core2 showed increases in T_m of 13°C and 5°C, respectively, over native met hG-CSF.

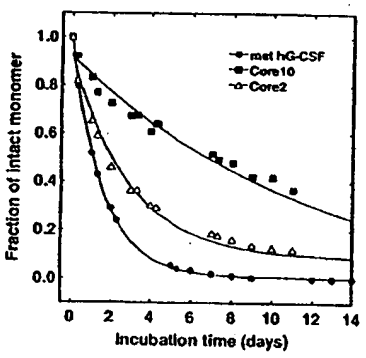


Fig. 5. Shelf life of hG-CSF analogs. Shelf life was estimated by incubating the proteins at elevated temperature (50°C) and using size exclusion chromatography to observe disappearance of monomeric protein. Rate constants were determined by a first order exponential fit of the fraction monomer remaining/time curves. Core2 and core10 showed five- and 10-fold improvements in storage stability, respectively, over met hG-CSF controls.

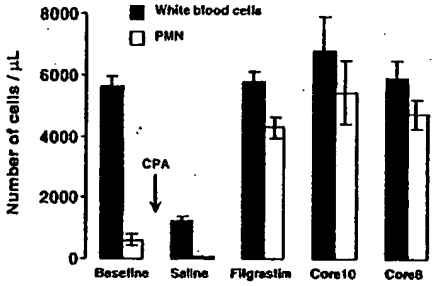


Fig. 6. In vivo granulopoietic activity of hG-CSF analogs. Mice were rendered neutropenic with a single intraperitoneal injection of 200 mg/kg cyclophosphamide (CPA). Beginning 24 h later and for 4 consecutive days, the mice were given a daily intravenous injection of 100 µg/kg of native hG-CSF (filgrastim, Amgen), an hG-CSF analog, or saline. On day 5, granulopoietic activity was determined by counting the number of white blood cells and polymorphonuclear neutrophils (PMN). The designed analogs (core8 and core10) were as effective as controls in eliciting a granulopoietic response.

active as wild-type hG-CSF (data not shown). The designed analogs were also as effective as wild type in increasing white blood cell and polymorphonuclear neutrophil levels in the neutropenic mouse (Fig. 6). Neutropenia, characterized by an abnormally low level of neutrophils in the blood, was induced by injection of cyclophosphamide. Reversal of this effect by the designed analogs shows that granulopoietic activity was also retained in vivo.

Pharmacokinetics

The pharmacokinetics of core10 and native hG-CSF (filgrastim, Amgen) was studied in cynomolgus monkeys after a single subcutaneous or intravenous injection of 5 µg/kg and after daily subcutaneous injections of 5 µg/kg for 28 d. Analysis of the serum concentration-time curves shows that subcutaneous injection of the designed analog results in greater systemic exposure (area under concentration-time curve, AUC) than the same dose of wild-type hG-CSF (Fig. 7B). This was true after a single dose on day 1 (78.8 vs. 54.6 ng-h/mL, data not shown), as well as on day 28 (37.2 vs. 17.4 ng-h/mL). There were no measurable differences in serum half-life. In the intravenous study, however, the half-life of core10 was three-fold shorter (1 vs. 3 h), and the AUC was significantly less (54.7 vs. 117.4 ng-h/mL), indicating that core10 is cleared faster (Fig. 7A). Taken together, these data indicate that the designed analog is absorbed more quickly from the subcutaneous compartment (absorption could not be measured directly given the small number of data points at early times). Improved absorption may be attributable to decreased aggregation or association of the designed protein. The increased monomer lifetime and decreased aggregation seen in our shelf-life studies and the improved thermal stability of the native conformation observed for core10 indicate a decrease in aggregation in the subcutaneous compartment. This possibility is supported by the fact that other protein therapeutics engineered for reduced aggregation also show faster absorption rates. For example, insulin Lispro and other rapid-acting insulin analogs that were designed to decrease their tendency to self-associate are absorbed faster than regular insulin after subcutaneous injection (Howey et al. 1994; Home et al. 1999).

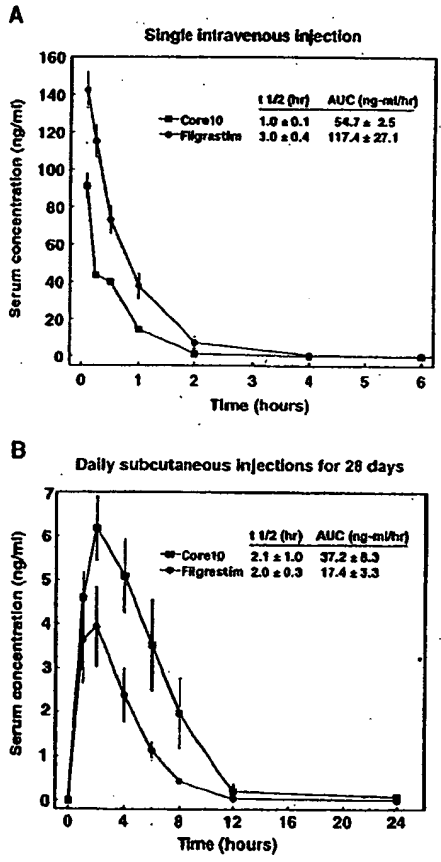


Fig. 7. Pharmacokinetics of hG-CSF analogs. Plasma concentrations of a designed hG-CSF analog or wild-type hG-CSF (filgrastim, Amgen) were determined after administration in cynomolgus monkeys. (*A*) Animals were given a single intravenous injection of 5 µg/kg or (*B*) daily subcutaneous injections of 5 µg/kg for 28 d. Noncompartmental analysis of the serum concentration-time curves shows that subcutaneous injections of the core10 analog resulted in greater systemic exposure (area under concentration-time curve, AUC) than the same dose of wild-type hG-CSF, whereas there was no change in serum half-life ($t_{1/2}$). In the intravenous study, the AUC was significantly less and the $t_{1/2}$ three-fold shorter, indicating that core10 was cleared faster.

Comparison to published G-CSF variants

In vitro and cassette mutagenesis studies have shown that alterations of the N-terminal region of G-CSF can lead to improved granulopoietic activity (Kuga et al. 1989; Okabe et al. 1990). Point mutations at Cys17 have also been found to affect shelf life; replacement with Ala led to an increase, Ser had no effect, and large residues (Ile, Tyr, Arg) led to a decrease (Ishikawa et al. 1992). In contrast, our core10 sequence, which has a large residue (Leu) at this position, showed an improved shelf life. This may be explained by the observation that in a Cys17Leu point mutant, Leu's side chain would clash with the aromatic ring of the nearby Phe at position 113. This steric clash does not occur in core10, however, because the Phe at 113 is replaced by Leu and, in compensation for this change, two nearby Leu's become Phe's (at positions 78 and 168). Thus, multiple mutations allow complementary repacking of the hydrophobic core in the core10 mutant and may be responsible for its enhanced stability and shelf life.

Significant improvements in thermal stability were also observed when the seven helical Gly residues in G-CSF were replaced with Ala to form point, double, and triple mutants (Bishop et al. 2001). Substitutions at positions 26, 28, 149, and 150 were the most effective. The investigators attributed the stabilizing effect to the enhancement in α-helical propensity associated with the Gly/Ala substitutions. These data support our suggestion that the heightened thermal stability seen with our mutants (which also contain a Gly/Ala substitution at position 28) is at least in part attributable to an improvement in helical propensity.

Probing the robustness of PDA with a homology modeled core position

As pointed out previously, the homology modeling of human G-CSF onto the bovine structure was straightforward for the most part because the replaced residues were primarily solvent exposed and no rearrangement of the backbone was necessary. The change at one core position, however, G167V, induced a steric clash and energy minimization of the entire protein was used to relieve the strain. We decided to assess the impact of this manipulation by doing an additional design (core167V) in which the variable residues were essentially the same as in the deep core design except that position 167 was also allowed to vary. We found that Val167 mutated to Ala (the other mutations were essentially the same as for core10). To probe the plasticity of the core, instead of using this PDA optimal sequence, which only had two mutations in this region, we ran experiments on another high-scoring sequence (core14_V167A) that had additional mutations (14 total, including L157I, F160W, and L161F). This sequence was chosen because it balanced an extensive number of mutations with a relatively high design score. Although it ranked 21st in the sequence energy list and was 2 kcal/mole less favorable than the optimal sequence, it was still biologically active and as stable as wild type (T_m of 61°C) (Figs. 2, 4). This indicates that optimization with PDA is fairly robust, and that the protein core can be quite plastic and can accommodate large changes without sacrificing stability or function.

Conclusions

PDA is a powerful ultrahigh throughput computational screening method. Its ability to screen up to 10^{80} sequences and allow multiple simultaneous mutations significantly increases the likelihood of finding new and improved proteins. In this study, PDA was used to develop improved analogs for a therapeutically important protein, hG-CSF. The novel proteins showed enhanced thermal stabilities and shelf life while retaining biological activity. Analysis of the mutants and results obtained with derived sequences indicates that the heightened stability is attributable to improvements in helical propensity and the elimination of a free cysteine; improved core packing and optimized hydrophobic burial of side chains may also be important. Pharmacokinetic studies indicate that subcutaneous injection of the most stable variant results in greater systemic exposure, probably attributable to improved absorption from the subcutaneous compartment.

These results show that PDA can be successfully applied to proteins of therapeutic interest. They also illustrate the value of its precise control over the site and type of mutations, allowing for the rational design of desired properties such as improved stability and pharmacokinetics and the elimination of undesirable ones such as toxicity and antigenicity. These features are particularly important in the design of therapeutic proteins. PDA thus has great potential as a powerful in silico tool for therapeutic protein design.

Materials and methods

Template structure preparation

The template structure for the designed proteins was produced by homology modeling using the crystal structure of bovine G-CSF (Brookhaven Protein Data Bank code 1bgc) as the starting point. The program BIOGRAF (Molecular Simulations Inc., San Diego, CA) was used to generate explicit hydrogens on the structure, which was then minimized for 50 steps using the conjugate gradient method and the Dreiding II force field (Mayo et al. 1990). The residues that differ in the bovine sequence or were not present in the bovine crystal structure were replaced with the human residues for those positions. The conformations of the replaced side chains were optimized using PDA (Dahiyat and Mayo 1997a,b), and the entire structure was minimized again for 50 steps. This minimized structure was used as the template for all the designs.

Protein design

Analogs of hG-CSF were designed by simultaneously optimizing residues in the buried core of the protein using PDA. The computational details, residue classification, potential functions, and parameters used for van der Waals interactions, solvation, and hydrogen bonding are described in previous work (Dahiyat and Mayo 1996, 1997a). An expanded version of the backbone-dependent rotamer library of Dunbrack and Karplus (Dunbrack and Karplus 1993) was used in all the calculations. The global optimum sequence from each design was selected for characterization and experimental testing, except for core167V in which the 21st ranked sequence was used. Calculations were generally performed overnight using 16 processors of an SGI Origin 2000 with 32 R10000 processors running at 195 MHz. The length of the runs varied from 1 to several hours of CPU time.

Cloning and expression

A gene for met hG-CSF was synthesized from partially overlapping oligonucleotides (~100 bases) that were extended and PCR amplified. Codon usage was optimized for *E. coli* and several restriction sites were incorporated to ease future cloning. These partial genes were cloned into a vector and transformed into *E. coli* for sequencing. Several of these gene fragments were then cloned into adjacent positions in an expression vector (pET17 or pET21) to form the full-length gene for met hG-CSF (528 bases) and transformed into *E. coli* for expression. Protein was expressed in *E. coli* in insoluble inclusion bodies and its identity was confirmed by immunoblot of SDS-PAGE using a commercial mAb against hG-CSF.

Refolding, purification, and storage

The protein inclusion bodies were solubilized in detergent and refolded in the presence of $CuSO_4$ to promote formation of native disulfide bonds (Lu et al. 1992). A size-exclusion column (10 mm × 300 mm loaded with Superdex prep 75 resin purchased from Pharmacia) was loaded with protein and eluted at a flow rate of 0.8 mL/min using the column buffer (100 mM Na_2SO_4, 50 mM Tris, pH 7.5). The peaks were monitored at dual wavelengths of 214 nm and 280 nm. Albumin, carbonic anhydrate, cytochrome C, and aprotinin were used to calibrate the molecular size of proteins versus elution time. The monomeric peak that elutes around the expected elution time for each protein was collected and the buffer was exchanged into 10 mM NaOAc at pH 4 for biophysical characterization. For long-term storage, a buffer of 5% sorbitol, 0.004% Tween 80, and 10 mM NaOAc at pH 4 was used. A pH of 4 was chosen for these buffers to be consistent with the commercial formulation of hG-CSF (Amgen), which was used as a control. The proteins were >98% pure as judged by reversed phase high performance liquid chromatography (HPLC) on a C4 column (3.9 mm × 150 mm) with a linear acetonitrile-water gradient containing 0.1% TFE. The identities of all proteins were confirmed by comparing the molecular mass measured by mass spectrometry with corresponding molecular mass calculated using the protein sequences.

Spectroscopic characterization

Protein samples were 50 μM in 50 mM sodium phosphate at pH 5.5. Concentrations were determined using UV spectrophotometry. Protein structure was assessed by CD. CD spectra were measured on an Aviv 202DS spectrometer equipped with a Peltier temperature control unit using a 1-mm path length cell. Thermal stability was assessed by monitoring the temperature dependence of the CD signal at 222 nm (Kolvenbach et al. 1997). A buffer of 10 mM NaOAc was used at pH 4.0 and data were collected every 2.5°C with an averaging time of 5 sec and an equilibration time of 3 min. Thermal denaturation curves were smoothed using KaleidaGraph. The melting temperature (T_m) of each protein was derived from the derivative curve of the ellipticity at 222 nm versus temperature. The T_m values were reproducible to within 2°C for the same protein at the concentrations used.

Storage stability

The storage stability of the designed proteins was assessed by incubation at both 37°C and 50°C under solution conditions identical to that used in the commercial formulation of hG-CSF (filgrastim, Amgen). Because aggregation and chemical degradation are the predominant mechanisms of inactivation of G-CSF (Herman et al. 1996), accelerated degradation was followed by observing the disappearance of monomeric protein with both size-exclusion and reverse-phase chromatography. Rate constants for shelf-life estimation were determined by a first-order exponential fit of the fraction monomer remaining/time curves using KaleidaGraph (Synergy Software).

Cell proliferation assay

Granulopoietic activity was measured by quantifying cell proliferation as a function of protein concentration using Ba/F3 (murine lymphoid) cells stably transfected with the gene encoding the human Class 1 G-CSF receptor (Avalos et al. 1995). Cell proliferation was detected by 5-bromo-2′-deoxyuridine (BrdU) incorporation quantified by a BrdU-specific ELISA kit (Boehringer Mannheim).

In vivo biological activity

Granulopoietic activity was determined in the neutropenic mouse (Hattori et al. 1990). C57BL/6 mice were rendered neutropenic with a single intraperitoneal injection of 200 mg/kg cyclophosphamide (CPA). Beginning 24 h later and for 4 consecutive days, the mice were given a daily intravenous injection of 100 μg/kg of an hG-CSF analog, met hG-CSF produced in our laboratory, clinically available hG-CSF (filgrastim, Amgen), or saline. On day 5, 6 h after the final dose, the animals were killed, blood samples were collected, and granulopoietic activity was determined by counting the number of white blood cells and polymorphonuclear neutrophils.

Pharmacokinetics

Plasma concentrations of a designed hG-CSF analog or wild-type hG-CSF (filgrastim, Amgen) were determined following administration in cynomolgus monkeys. Animals were given a single intravenous injection of 5 μg/kg or daily subcutaneous injections of 5 μg/kg for 28 d. In the intravenous study, blood samples were collected at 0 (predose), 5, 15, and 30 min and 1, 2, 4, 6, 8, 12, and 24 h postdosing. In the subcutaneous studies, blood samples were collected at 0 (predose), 1, 2, 4, 6, 8, 12, and 24 h postdosing on day 1 and day 28. All samples were immediately placed on wet ice and centrifuged at 28°C. The resultant plasma was then frozen and

stored (−70°C). Plasma concentrations were determined using an enzyme-linked immunosorbent assay (Quantikine human G-CSF ELISA, R&D Systems, Minneapolis, MN), performed per manufacturers instructions except that samples were diluted in PBS, 5% nonfat dry milk, and 0.05% Tween 20, and the incubation was extended to overnight at 4°C. Plasma concentrations of the designed hG-CSF analog and filgrastim were estimated from their corresponding standard curves. Pharmacokinetic parameters were calculated by noncompartmental analysis. The terminal slope (λz) was estimated by linear regression through the last time points of the log concentration versus time curves and used to calculate the terminal half-life ($t_{1/2}$). The area under the curve from time of dosing through the last time point (AUC_{0-t}) was calculated by the linear trapezoid method.

Acknowledgments

We thank Dr. Belinda Avalos (Ohio State University) for kindly supplying the Ba/F3 cell line transfected with the hG-CSF receptor. We also thank Dr. Steven Adams (American College of Laboratory Animal Medicine) and LAB Preclinical Research Institute Inc., (Quebec, Canada) for conducting the monkey studies.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Aritomi, M., Kunishima, N., Okamoto, T., Kuroki, R., Ota, Y., and Morikawa, K. 1999. Atomic structure of the GCSF-receptor complex showing a new cytokine-receptor recognition scheme. *Nature* 401: 713–717.

Avalos, B.R., Hunter, M.G., Parker, J.M., Ceselski, S.K., Druker, B.J., Corey, S.J., and Mehta, V.B. 1995. Point mutations in the conserved box 1 region inactivate the human granulocyte colony-stimulating factor receptor for growth signal transduction and tyrosine phosphorylation of p75c-rel. *Blood* 85: 3117–3126.

Bishop, B., Koay, D.C., Sartorelli, A.C., and Regan, L. 2001. Reengineering granulocyte colony-stimulating factor for enhanced stability. *J. Biol. Chem.* 276: 33465–33470.

Bolon, D.N. and Mayo, S.L. 2001. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* 98: 14274–14279.

Chang, C.C., Chen, T.T., Cox, B.W., Dawes, G.N., Stemmer, W.P., Punnonen, J., and Patten, P.A. 1999. Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* 17: 793–797.

Chen, K.Q. and Arnold, F.H. 1991. Enzyme engineering for nonaqueous solvents: Random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnology* 9: 1073–1077.

Crameri, A., Whitehorn, E.A., Tate, E., and Stemmer, W.P. 1996. Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat. Biotechnol.* 14: 315–319.

Dahiyat, B.I. and Mayo, S.L. 1996. Protein design automation. *Protein Sci.* 5: 895–903.

———. 1997a. De novo protein design: Fully automated sequence selection. *Science* 278: 82–87.

———. 1997b. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.* 94: 10172–10177.

Desjarlais, J.R. and Handel, T.M. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci.* 4: 2006–2018.

Dunbrack, R.L. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins—an application to side-chain prediction. *J. Mol. Biol.* 230: 543–574.

Grossmann, M., Leitolf, H., Weintraub, B.D., and Szkudlinski, M.W. 1998. A rational design strategy for protein hormone superagonists. *Nat. Biotechnol.* 16: 871–875.

Hattori, K., Shimizu, K., Takahashi, M., Tamura, M., Oheda, M., Ohsawa, N., and Ono, M. 1990. Quantitative in vivo assay of human granulocyte colony-stimulating factor using cyclophosphamide-induced neutropenic mice. *Blood* 75: 1228–1233.

Heikoop, J.C., van den Boogaart, P., Mulders, J.W., and Grootenhuis, P.D. 1997. Structure-based design and protein engineering of intersubunit disulfide bonds in gonadotropins. *Nat. Biotechnol.* 15: 658–662.

Hellinga, H.W. and Richards, F.M. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci.* 91: 5803–5807.

Herman, A.C., Boone, T.C., and Lu, H.S. 1996. Characterization, formulation, and stability of Neupogen (Filgrastim), a recombinant human granulocyte-colony stimulating factor. *Pharm. Biotechnol.* 9: 303–328.

Hill, C.P., Osslund, T.D., and Eisenberg, D. 1993. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc. Natl. Acad. Sci.* 90: 5167–5171.

Home, P.D., Barriocanal, L., and Lindholm, A. 1999. Comparative pharmacokinetics and pharmacodynamics of the novel rapid-acting insulin analogue, insulin aspart, in healthy volunteers. *Eur. J. Clin. Pharmacol.* 55: 199–203.

Horan, T., Wen, J., Narhi, L., Parker, V., Garcia, A., Arakawa, T., and Philo, J. 1996. Dimerization of the extracellular domain of granulocyte-colony stimulating factor receptor by ligand binding: A monovalent ligand induces 2:2 complexes. *Biochemistry* 35: 4886–4896.

Howey, D.C., Bowsher, R.R., Brunelle, R.L., and Woodworth, J.R. 1994. [Lys(B28), Pro(B29)]-human insulin. A rapidly absorbed analogue of human insulin. *Diabetes* 43: 396–402.

Ishikawa, M., Iijima, H., Satake-Ishikawa, R., Tsumura, H., Iwamatsu, A., Kadoya, T., Shimada, Y., Fukamachi, H., Kobayashi, K., Matsuki, S., et al. 1992. The substitution of cysteine 17 of recombinant human G-CSF with alanine greatly enhanced its stability. *Cell Struct. Funct.* 17: 61–65.

Jiang, X., Farid, H., Pistor, E., and Farid, R.S. 2000. A new approach to the design of uniquely folded thermally stable proteins. *Protein Sci.* 9: 403–416.

Kolvenbach, C.G., Narhi, L.O., Philo, J.S., Li, T., Zhang, M., and Arakawa, T. 1997. Granulocyte-colony stimulating factor maintains a thermally stable, compact, partially folded structure at pH2. *J. Pept. Res.* 50: 310–318.

Kraemer-Pecore, C.M., Wollacott, A.M., and Desjarlais, J.R. 2001. Computational protein design. *Curr. Opin. Chem. Biol.* 5: 690–695.

Kuga, T., Komatsu, Y., Yamasaki, M., Sekine, S., Miyaji, H., Nishi, T., Sato, M., Yokoo, Y., Asano, M., Okabe, M., et al. 1989. Mutagenesis of human granulocyte colony stimulating factor. *Biochem. Biophys. Res. Commun.* 159: 103–111.

Lovejoy, B., Cascio, D., and Eisenberg, D. 1993. Crystal structure of canine and bovine granulocyte-colony stimulating factor (G-CSF). *J. Mol. Biol.* 234: 640–653.

Lowman, H.B. and Wells, J.A. 1993. Affinity maturation of human growth hormone by monovalent phage display. *J. Mol. Biol.* 234: 564–578.

Lu, H.S., Clogston, C.L., Narhi, L.O., Merewether, L.A., Pearl, W.R., and Boone, T.C. 1992. Folding and oxidation of recombinant human granulocyte colony stimulating factor produced in *Escherichia coli*. Characterization of the disulfide-reduced intermediates and cysteine—serine analogs. *J. Biol. Chem.* 267: 8770–8777.

Malakauskas, S. and Mayo, S. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* 5: 470–475.

Marshall, S.A. and Mayo, S.L. 2001. Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* 305: 619–631.

Mayo, S.L., Olafson, B.D., and Goddard III, W.A. 1990. Dreiding: A generic forcefield for molecular simulations. *J. Phys. Chem.* 94: 8897–8909.

Okabe, M., Asano, M., Kuga, T., Komatsu, Y., Yamasaki, M., Yokoo, Y., Itoh, S., Morimoto, M., and Oka, T. 1990. In vitro and in vivo hematopoietic effect of mutant human granulocyte colony-stimulating factor. *Blood* 75: 1788–1793.

Pokala, N. and Handel, T.M. 2001. Review: Protein design—where we were, where we are, where we're going. *J. Struct. Biol.* 134: 269–281.

Reidhaar-Olson, J.F., De Souza-Hart, J.A., and Selick, H.E. 1996. Identification of residues critical to the activity of human granulocyte colony-stimulating factor. *Biochemistry* 35: 9034–9041.

Shimaoka, M., Shifman, J.M., Jing, H., Takagi, J., Mayo, S.L., and Springer, T.A. 2000. Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* 7: 674–678.

Stemmer, W.P. 1994. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 370: 389–391.

Street, A.G. and Mayo, S.L. 1999. Computational protein design. *Structure Fold. Des.* 7: R105–109.

Strop, P. and Mayo, S.L. 1999. Rubredoxin variant folds without iron. *J. Am. Chem. Soc.* 121: 2341–2345.

Young, D.C., Zhan, H., Cheng, Q.L., Hou, J., and Matthews, D.J. 1997. Characterization of the receptor binding determinants of granulocyte colony stimulating factor. *Protein Sci.* 6: 1228–1236.

Zhao, H., Giver, L., Shao, Z., Affholter, J.A., and Arnold, F.H. 1998. Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nat. Biotechnol.* 16: 258–261.

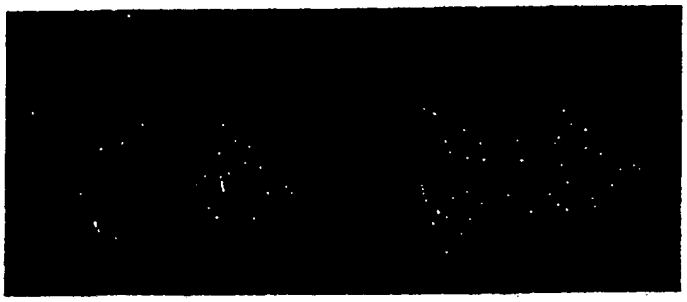
Proteins from Scratch

William F. DeGrado

Not long ago, it seemed inconceivable that proteins could be designed from scratch. Because each protein sequence has an astronomical number of potential conformations, it appeared that only an experimentalist with the evolutionary life span of Mother Nature could design a sequence capable of folding into a single, well-defined three-dimensional structure. But now, on page 82 of this issue, Dahiyat and Mayo (*1*) describe a new approach that makes de novo protein design as easy as running a computer program. Well almost...

The intellectual roots of this new work go back to the early 1980s when protein engineers first thought about designing proteins (*2*). At that point, the prediction of a protein's three-dimensional structure from its sequence alone seemed a difficult proposition. However, they opined that the inverse problem—designing an amino acid sequence capable of assuming a desired three-dimensional structure—would be a more tractable problem, because one could "over-engineer" the system to favor the desired folding pattern. Thus, the problem of de novo protein design reduced to two steps: selecting a desired tertiary structure and finding a sequence that would stabilize this fold. Dahiyat and Mayo have now mastered the second step with spectacular success. They have distilled the rules, insights, and paradigms gleaned from two decades of experiments (*3*) into a single computational algorithm that predicts an optimal sequence for a given fold. Further, when put to the test the algorithm actually predicted a sequence that folded into the desired three-dimensional structure. Thus, the rules of protein folding and computational methods for de novo design may now be sufficiently defined to allow the engineering of a variety of proteins.

Dahiyat and Mayo's program divides the interactions that stabilize protein structures into three categories: interactions of side chains that are exposed to solvent, of side chains buried in the protein interior, and of parts of the protein that occupy more interfacial positions. Exposed residues contribute to stability, primarily through conformational preferences and weakly attractive, solvent-exposed polar interactions (*4*). The burial of hydrophobic residues in the well-packed interior of a protein provides an even more powerful driving force for folding. The side chains in the interior of a protein adopt unique conformations, the prediction of which is a large combinatorial problem.



Better than the real thing. The natural zinc finger protein Zif268 (left) is stabilized in part by a core of hydrophobic (green) side chains and metal-chelating side chains (red). In the designed protein FSD-1 (right), the Zif268 core is retained but the metal-chelating His residues and one of the Cys residues of Zif268 are converted to hydrophobic Phe and Ala residues, thereby extending the hydrophobic core. The fourth metal ligand Cys^6 is converted to a Lys residue. The apolar portion of this interfacial residue shields the hydrophobic core, whereas its ammonium group is exposed to solvent. The helix is also stabilized by an N-capping interaction (*19*), which presumably also stabilizes the structure.

One important simplifying assumption arose from the early work of Jainin *et al.* (*5*), who showed that each individual side chain can adopt a limited number of low-energy conformations (named rotamers), reducing the number of probable conformers available to a protein. This work was subsequently extended to the design of proteins containing only the most favorable rotamers (*6*). Although the side chains in natural proteins deviate from ideality in a few cases (complicating the prediction of the structures of natural proteins), these deviations need not be considered in the design of idealized proteins. Thus, various algorithms have been developed to examine all possible hydrophobic residues in all possible rotameric states, to find combinations that efficiently fill the interior of a protein. A complementary approach uses genetic methods to exhaustively search for sequences capable of filling a protein core (*7*), and this work has been adapted for the de novo design of proteins (*8*).

Interfacial residues are also quite important for protein stability (*9*, *10*). They are often amphiphilic (for example, Lys, Arg, and Tyr) and their apolar atoms can cap the hydrophobic core, while their polar groups engage in electrostatic and hydrogen-bonded interactions.

Until recently, protein designers have frequently concentrated on quantifying the energetics associated with just one of these three types of interactions (*3*). However, de novo design is best approached by simultaneously considering all of the side chains in the protein—unfortunately, a very high-order combinatorial problem. For instance, the volume available to the interior side chains depends on the nature and conformation of the residues at the interfacial positions and vice versa. Dahiyat and Mayo assumed that each of these three features had been adequately quantitated to provide a useful empirical energy function for protein design. Their program combines a number of feaures taken from earlier potential functions and includes a penalty for exposing hydrophobic groups to solvent. Another essential innovation included in their program is an implementation of the Dead-End Elimination theorem, to efficiently search through sequence and side chain rotamer space.

Dahiyat and Mayo's target fold is a zinc finger, a motif with a well-established history in protein structure prediction and design. In an early, prescient paper, Berg correctly inferred that this His^2Cys^2 Zn-binding motif must feature a β-β-α fold that would position the ligating groups in a tetrahedral array around the bound Zn(II) (*11*). Favorable metal ion–ligand interactions together with a small apolar core help stabilize the three-dimensional structure of this compact fold. More recently, Imperiali and co-workers have designed a peptide that folded into this motif, even in the absence of metal ions (*12*). The design included a D-amino acid to stabilize a type II′ turn, and a large, rigid tricyclic side chain that may help consolidate the hydrophobic core. This work was particularly ex-

An enhanced version of this Perspective with links to additional resources is available for Science Online subscribers at www.sciencemag.org

The author is in the Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6059, USA. E-mail: wdegrado@mail.med.upenn.edu

citing because, before their studies, it was not expected that sequences as short as 25 residues in length could fold into stable tertiary structures.

Now, Dahiyat and Mayo take these studies one step further through the design of a sequence composed of only natural amino acids that adopts the zinc finger motif. As input to their program, they introduced the coordinates of the backbone atoms from the crystal structure of the second domain of the zinc finger protein Zif268. The program then evaluated a total of 10^{62} possible side chain–rotamer combinations to find a sequence capable of stabilizing this fold without a bound metal ion. The resulting protein sequence shares a small hydrophobic core with its predecessor from Zif268. However, in the newly designed protein FSD-1 the core is enlarged through the addition of hydrophobic residues that fill the space vacated by the removal of the metal-binding site (see the figure). This increase in the size of the hydrophobic core together with the enhancements in the propensity for forming the appropriate secondary structure provide an adequate driving force for folding. The designed miniprotein actually folds into the desired structure as assessed by nuclear magnetic resonance spectroscopy, and the observed structure closely resembles the three-dimensional structure of Zif268.

Because of its small size, the protein is marginally stable. A Van't Hoff analysis of the thermal unfolding curve gives a change in the enthalpy (ΔH_{vH}) of approximately −10 kcal/mol, and indicates that the protein is about 90 to 95% folded at low temperatures (*13*). The small value ΔH_{vH} and the lack of strong cooperativity in the unfolding transition are expected for a native-like protein of this very small size (*14*). Thus, FSD-1 is the smallest protein known to be capable of folding into a unique structure without the thermodynamic assistance of disulfides, metal ions, or other subunits. This important accomplishment illustrates the impressive ability of Dahiyat and Mayo's program to design highly optimized sequences.

This new achievement caps a banner year for de novo protein design. Earlier, Regan (*15*) answered the challenge of changing a protein's tertiary structure by altering no more than 50% of its sequence. And although Dahiyat and Mayo have demonstrated that the stabilizing metal-binding site is not necessary in their system, Caradonna, Hellinga, and co-workers (*16*) have made impressive progress in automating the introduction of functional metal-binding sites into the three-dimensional structures of natural proteins. Further, other workers (*17*) have used less automated approaches to successfully introduce functionally and spectroscopically interesting metal-binding sites into de novo designed proteins.

To date, the most computationally intensive protein design problems have been the redesign of natural proteins of known three-dimensional structure. But the new automated approaches open the door to the de novo design of structures with entirely novel backbone conformations. It will be interesting to see if Dahiyat and Mayo's approach of designing an optimal sequence for a given fold is sufficient, or if it will be necessary also to destabilize alternate possible folds. Indeed, when using an earlier version of their algorithm to repack the interior of the coiled coil from GCN4, they had to retain the identity of a buried Asn residue from the wild-type protein. Although the inclusion of this Asn actually destabilized the desired fold, it was nevertheless essential to avoid the formation of alternate, unwanted conformers (*18*). The ability to ask such focused questions will reveal much about how natural proteins adopt their folded conformations while simultaneously allowing the design of entirely new polymers for applications ranging from catalysis to pharmaceuticals.

References and Notes

1. B. I. Dahiyat and S. L. Mayo, *Science* **278**, 82.
2. K. E. Drexler, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5275 (1981); C. Pabo, *Nature* **301**, 200 (1983).
3. W. F. DeGrado, Z. R. Wasserman, J. D. Lear, *Science* **243**, 622 (1989); J. W. Bryson *et al.*, *ibid.* **270**, 935 (1995); M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Opin. Struct. Biol.* **6**, 3 (1996).
4. R. Munoz and L. Serrano, *Proteins* **20**, 301 (1994); C. A. Kim and J. M. Berg, *Nature* **362**, 267 (1993); D. L. Minor and P. S. Kim, *ibid.* **367**, 660 (1994); C. K. Smith, J. M. Withka, L. Regan, *Biochemistry* **33**, 5510 (1994).
5. J. Janin, S. Wodak, M. Levitt, B. Maigret, *J. Mol. Biol.* **125**, 37 (1978).
6. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987); J. R. Desjarlais and T. M. Handel, *Protein Sci.* **4**, 2006 (1995); X. Jing, E. J. Bishop, R. S. Farid, *J. Am. Chem. Soc.* **119**, 838 (1997).
7. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).
8. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *ibid.* **262**, 1680 (1993).
9. K. J. Lumb and P. S. Kim, *ibid.* **271**, 1137 (1996); Y. Yu, O. D. Monera, R. S. Hodges, P. L. Privalov, *J. Mol. Biol.* **255**, 367, (1996).
10. A. C. Braisted and J. A. Wells, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5688 (1996).
11. J. M. Berg, *Ibid.* **65**, 99 (1988).
12. M. D. Struthers, R. P. Cheng, B. Imperiali, *Science* **271**, 342 (1996).
13. This Van't Hoff analysis of the protein is approximate because of the lack of definition of the pre- and posttransition baselines.
14. P. Alexander, S. Fahnestock, T. Lee, J. Orban, P. Bryn, *Biochemistry* **31**, 3597 (1992).
15. S. Dalal, S. Balasubramanian, L. Regan, *Nat. Struct. Biol.* **4**, 548 (1997).
16. A Pinto, H. W. Hellinga, J. P. Caradonna, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5562 (1997); C. Coldren, H. W. Hellinga, J. P. Caradonna, *Ibid.*, p. 6635.
17. B. R. Gibney, S. E. Mutholland, F. Rabanal, P. L. Dutton, *ibid.* **93**, 15041 (1996); M. P. Scott, J. Biggins, *Protein Sci.* **6**, 340 (1997); P. A. Arnold, W. R. Shelton, D. R. Benson, *J. Am. Chem. Soc.* **119**, 3181 (1997); G. R. Dieckman *et al.*, *Ibid.*, p. 6195.
18. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401 (1993); K. J. Lumb and P. S. Kim, *Biochemistry* **34**, 8642 (1995).
19. L. G. Presta and G. D. Rose, *Science* **240**, 1632 (1988); J. S. Richardson and D. C. Richardson, *Ibid.*, p. 1648.

Combinatorial protein design

Jeffery G Saven

Combinatorial protein libraries permit the examination of a wide range of sequences. Such methods are being used for *de novo* design and to investigate the determinants of protein folding. The exponentially large number of possible sequences, however, necessitates restrictions on the diversity of sequences in a combinatorial library. Recently, progress has been made in developing theoretical tools to bias and characterize the ensemble of sequences that fold into a given structure – tools that can be applied to the design and interpretation of combinatorial experiments.

Addresses
Department of Chemistry, University of Pennsylvania, 231 South 34 Street, Philadelphia, Pennsylvania 19104, USA;
e-mail: saven@sas.upenn.edu

Current Opinion in Structural Biology 2002, 12:453–458

0959-440X/02/$ – see front matter

Introduction

The discovery and design of novel proteins can lead to new, potentially practical proteins and can also enhance our understanding of protein biochemistry. Designing well-structured, soluble proteins is difficult, however, because of their complexity. Such proteins are large (tens to hundreds of amino acid residues) and have many variables that specify the folded state, including sequence, backbone topology and sidechain conformation. Design involves identifying those sequences that fold into a given structure from a huge ensemble of possible sequences. This search is aided, in part, by the large degree of consistency seen in folded proteins. On average, a folded structure is well packed, hydrophobic residues are sequestered from solvent and most potential hydrogen bond interactions are satisfied. This consistency, however, is often complex, may have little simplifying symmetry and involves predominantly noncovalent interactions. Such interactions are some of the most difficult to accurately quantify. As such, estimating the free energies associated with mutation or structural ordering remains a subtle area of computational research. Nonetheless, many molecular potentials do contain a 'best parameterization' of many of the interatomic interactions and forces that we know are important for stabilizing proteins. In some cases, such potentials have been used with striking success in protein design [1••]. Given that these potentials are necessarily approximate, however, one promising approach is to use the partial information contained in these functions in a probabilistic manner. A probabilistic or statistical approach is also appropriate for characterizing the full variability of sequences that fold to a common structure, because there are likely to be an enormous number of such sequences. Such statistical methods can be applied in 'shotgun' approaches to *de novo* protein design. Combinatorial experiments create and assay many sequences in order to overcome shortcomings in our understanding of folding or other molecular properties. Even though combinatorial methods can address large numbers of sequences (10^4–10^{12}), these numbers are still infinitesimal in comparison to the numbers of possible sequences (e.g. $20^{100} \approx 10^{130}$ for a 100-residue protein). Thus, methods for winnowing and focusing sequence space are a vital component of combinatorial protein design. Herein, I briefly discuss combinatorial methods for full sequence design. I also review recent theoretical developments in characterizing sequence ensembles — developments that can be applied to the design and interpretation of combinatorial experiments.

Directed protein design

There has been much effort — and success — in developing computational methods for 'directed' protein design. By 'directed protein design', I mean the identification of a sequence (or a small set of sequences) that is likely to fold into a predetermined backbone structure. Each such sequence can then be synthesized to confirm its folded structure and other molecular properties. Early efforts in design identified proteins with substantial order, but not necessarily well-defined tertiary structure [2]. Because an enormous number of sequences are possible even for small proteins (<50 residues), computational methods have dramatically accelerated successful design. Typically, such methods are implemented as an optimization process, whereby amino acid identity and sidechain conformation are varied in order to optimize a scoring function that quantifies sequence/structure compatibility. Exhaustive searching of all m^N possible sequences (where m is the number of different amino acid types or 'states' per residue and N is the number of residues in a target protein structure) is feasible only if a small number of residues N are allowed to vary or if the number of amino acids m is greatly reduced. If, in the optimization process, the different sidechain conformations (rotamer states) of each amino acid are also considered (see [3]), the complexity of the search increases still further, because m, the number of possible 'states' per residue, increases by a factor of ten or more. Although complete enumeration is typically not feasible, sequence space can be sampled in a directed manner in order to find optimal (or nearly optimal) sequences. Stochastic methods, such as genetic algorithms or simulated annealing, involve searching sequence space in a partially random fashion; on average, the search progressively moves toward better scoring (lower energy) sequences [4,5]. The partially random nature of the search permits escape from local minima in the sequence/rotamer landscape. Using a simplified model, the Takada and Tamura groups have included information about unfolded structures (negative design) in a stochastic search for a sequence with a 'funneled conformational energy landscape' [6]. One

47-residue three-helix bundle protein so selected has CD and NMR spectral features of folded proteins (W Jin, O Kambara, H Sasakawa, A Tamura, S Takada, personal communication). When applied to atomically detailed representations, the stochastic methods focus primarily on repacking the interior of a structure with hydrophobic residues [7] and have been applied to the wild-type structures of 434 Cro [8], ubiquitin [9], the B1 domain of protein G [10•], the WW domain [1••] and helical bundles [11,12]. Although, in many cases, these methods have identified experimentally viable sequences [1••,13], stochastic search methods need not identify global optima [14•]. For potentials comprising only site and pair interactions, elimination methods such as 'dead end elimination' can find the global optimum [14•,15–17]. Such methods successively remove individual amino acid rotamer states that cannot be part of the global optimum until no further states can be eliminated. The Mayo group applied such methods to automate the full sequence design of both a 28-residue zinc finger mimic [18] and, after predetermining hydrophobic and polar sites, a 51-residue homeodomain motif [19•]. The group has also redesigned portions of a variety of proteins [20–22]. Functional properties such as metal binding or catalysis may also be included as elements of the design process [23,24•]. The elements and algorithms of directed protein design have been the subject of several recent reviews [1••,25,26•].

Despite some striking successes, computational methods for directed design have limitations with respect to both identifying folding sequences and characterizing the features of protein sequences that share a common structure. Stochastic methods, such as simulated annealing or genetic algorithms, can be applied to large proteins and permit many sites to be varied simultaneously, but the computational times and resources required for such calculations are extensive, even for small proteins. When used as optimization methods, directed approaches will necessarily be sensitive to the energy or scoring function used. All energy functions in use in protein design, however, are necessarily approximate and uncertainties in the energy function may not merit the search for global optima. Furthermore, many naturally occurring proteins are not optimized. In fact, most proteins are only marginally stable (e.g. ΔG^{o} <10 kcal/mol for folding) [27]. In addition, sequences that function, for example, those that bind another molecule, need not be the global optimum with respect to structural stability. Although stochastic methods can sample such suboptimal sequences, in general an exponentially large number of them will be possible and such sampling will be time consuming. Thus, it is important to develop methods complementary to those used for directed protein design — methods that reveal the features of sequences that are likely to fold into a particular structure but that may not be structurally 'optimal'. Such computational methods will have application to a new class of protein design studies, combinatorial experiments, in which large numbers of proteins may be simultaneously synthesized and screened.

Combinatorial design

Combinatorial design provides a complementary approach to directed design for understanding sequence/ structure compatibility and discovering novel sequences that fold into a specific structure. Combinatorial methods are powerful tools for cases in which we have an incomplete understanding of molecular properties. In protein combinatorial design experiments, large numbers of sequences (libraries) are screened for evidence of folding into a predetermined structure. A combinatorial experiment has two key elements: creating a library with a desired degree of diversity and assaying for sequences with 'protein-like' properties in terms of their structure or function. Depending upon how the diversity is generated and assayed, experiments of this type can explore a large number of sequences, up to 10^{12} [28•]. Certainly, such methods can be used to discover 'hits', that is, a few sequences that are especially stable or that are unusually strong in their function or binding properties. In addition, combinatorial experiments readily generate a sequence ensemble. Thus, using combinatorial experiments, we can potentially 'expand the protein sequence database' and the diversity of these additional sequences will be at the control of the researcher. Features important to folding (and other properties) may be explored in a way that is decoupled from the evolutionary requirements of nature's proteins. For example, these methods have been used to identify helical proteins [29–31], ubiquitin variants [32], self-assembled protein monolayers [33], proteins with amyloid-like properties [33], metal-binding peptides [34] and stable interhelical oligomers [35]. Several excellent reviews of combinatorial experiments have appeared recently [36,37,38•,39••].

The complexity of combinatorial experiments implies that limitations must be placed on the sequences, because the number that can be created and screened (10^{6}–10^{12}) is infinitesimal compared to the number possible (e.g. 10^{130}). Limitations on sequence properties are often guided by qualitative chemical considerations, but quantitative computational methods will be helpful in designing and interpreting combinatorial experiments.

The Hecht group has probed the extent to which the patterning of hydrophobic and hydrophilic residues can successfully reduce complexity in combinatorial design. While maintaining the periodicity of α helices and β sheets in particular tertiary structures, such patterning is applied in order to expose hydrophilic residues to solvent and to sequester hydrophobic residues in the interior of the protein. Early targets were helical proteins; a fiducial 74-residue four-helix bundle was the template structure [40]. Such a structure has more than $20^{74} \approx 10^{96}$ possible sequences. After binary patterning, five hydrophobic and six hydrophilic amino acids were permitted at 24 interior and 36 exterior positions, respectively, thus reducing the total number of possible sequences to 10^{41}. From a protein library consistent with this binary patterning, a set of 50 correctly expressed sequences was selected for further

study. Around half of the 50 sequences isolated are protein-like in many respects [30], including their thermal denaturation [41]. About half the isolated sequences also bind heme [29] and many of these display carbon monoxide binding [42•] or peroxidase activity [43]. This is surprising given that such functions were not part of the design or selection of the sequences. In a second-generation design, the group added six residues to each of the four helices of one of the most protein-like sequences. The additional residues were combinatorially patterned, as in the original experiment [39••]. For these 102-residue sequences, the free energies of folding are increased 2–3-fold and the NMR data suggest well-determined structures. Using binary patterning of hydrophobicity consistent with an amphilic β sheet [44], the Hecht group has also identified proteins that aggregate to form amyloid fibrils [45] and crafted monomeric β proteins by introducing a nonpolar lysine mutation at the 'edge' strand of the target β sheet [46••].

Despite the striking results from hydrophobic patterning, more detailed methods for library design are merited. Many of the hydrophobically patterned sequences that appear well structured are not sufficiently soluble for NMR structure determination [46••] and, as a result, little is known concerning their structures at the atomic scale. Not all of the α-helical sequences exhibit the sharp thermal transition seen in natural proteins (usually associated with a large ΔH of folding). Such sequences may not possess well-packed interiors [41]. In natural proteins, the side-chains of most interior residues are well determined, as opposed to the variability that is obtained using hydrophobic patterning alone and that is observed in many *de novo* designed proteins [13,18]. A more fine-grained dictation of the amino acid identities is probably necessary for obtaining libraries that are rich in sequences with well-defined structures. Moreover, a more detailed specification of amino acid identities yields fewer sequences than hydrophobic patterning alone and further reduces the complexity of the library.

Theories of combinatorial libraries

Surveying the complete sequence landscape of proteins seems, at first glance, intractable to both experiment and computation. In addition to the enormous number of possible sequences, many examples exist in nature of dissimilar sequences folding to essentially the same structure. Hence, sequence properties are nontrivial and proteins sharing a common structure can be nonlocal in sequence space. Nonetheless, computational methods permit us to estimate the properties, particularly the amino acid probabilities, of sequences consistent with a target structure.

Repeated use of directed search methods can estimate the properties of an ensemble of sequences. Desjarlais and co-workers have used independent runs of their sequence prediction algorithm across an ensemble of closely related structures all consistent with a particular fold (JR Desjarlais *et al.*, personal communication). For each structure, an optimal 'nucleating' sequence is identified and subsequently the sequence/rotamer variability is explored throughout the structure. The method identifies effective reduced partition sums for each sequence/rotamer state and amino acid probabilities may be obtained at each residue position. The number of sequences decreases with stability, so the degree of complexity can be tuned by varying a cutoff in the effective free energies of the sequences. The method has been used to identify sequences consistent with the fold of a WW domain, a small β-sheet protein [1••], some of which are currently being experimentally characterized.

The amino acid frequencies can also be determined directly, using a statistical theory of combinatorial libraries [47,48••,49••]. Ideas from statistical mechanics are used to address the number and composition of sequences that are consistent with a particular backbone structure. The theory addresses the whole space of available compositions, not just the small fraction that is accessible to experiment and to computational enumeration and sampling. The theory takes as input a target backbone structure and a scoring or energy function for quantifying sequence/structure compatibility. Global and local features can be prespecified using constraints on the sequences. For example, such constraints can be used to determine the energy the sequences assume in the target structure, the patterning of amino acids and the number of each amino acid present (composition). The theory yields estimates of both the number of sequences consistent with these constraints and the amino acid probabilities at each residue position. These residue-specific probabilities are the most probable such set and are determined — as in statistical mechanics — by maximizing an effective entropy, whereby this maximization is subject to constraints. Just as in thermodynamics, the judicious use of constraints can be used to reduce the entropy or the number of possible sequences. Thus, these methods provide a systematic means to focus the library, winnowing numbers such as 10^{130} to numbers that are experimentally manageable, for example, 10^6. The theory agrees well with exact results obtained with lattice models of proteins [47,48••]. This method has been extended to realistic representations of proteins, in which the effects of sidechain packing are included in an atom-based manner [49••]. The calculated sequence probabilities of the immunoglobulin light chain binding domain of protein L are in agreement with the frequencies observed in combinatorial phage display experiments [50,51]. These statistical methods have several advantages. They may be applied to much larger proteins ($N > 100$ residues) and permit much larger sequence variation than many directed methods. They are sufficiently rapid that many backbone structures may be considered and those features that are robust with respect to minor structure modifications may be identified. Importantly, such methods provide perhaps the most natural input for a combinatorial experiment, the probabilities of the amino acids at each position among the sequences of a library. These amino acid

probabilities can also be used to identify specific amino acid sequences, which can then be synthesized; a consensus sequence comprising the most probable amino acid at each site can be selected or the probabilities can be used to bias a stochastic search for viable sequences (J Zou, JG Saven, unpublished data).

If the energy of the target state is one of the constraints, the statistical method reduces to an effective mean field theory. Mean field theories have seen extensive application in physical science and in biomolecular theory [52], and to protein evolution and natural sequence variability ([53]; H Kono, JG Saven, unpublished data). Voigt *et al.* [14•] have compared mean field theories with directed search methods for identifying ground state sequence/rotamer combinations in protein design. They found that, although often more rapid, mean field theories do not always identify such ground states. Interestingly, Voigt *et al.* applied the mean field theory to large proteins (subtilisin E and T4 lysozyme) to determine local site entropies, s_i, where $\exp(s_i)$ quantifies the effective number of amino acids allowed at residue i in a structure [54••,55]. Sites with large values of s_i, those most tolerant to mutation [56], are likely to support substitutions that improve stability or function when *in vitro* evolution experiments are used to explore sequence space [37]. For such experiments, the mutation rate is low enough that multiple mutations of strongly interacting sites are rare. Thus, mutations that improve 'fitness' are most likely to accumulate at sites that are the most 'decoupled' from other sites. Such mutations can potentially be targeted for variation in an *in vitro* evolution experiment.

Conclusions

Much recent progress has been seen in the design and discovery of new proteins, and combinatorial approaches are accelerating the pace. Such methods are most useful when our quantitative understanding of important protein properties, such as stability and catalytic activity, is limited. Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but, with combinatorial experiments, an enormous diversity of sequences can be generated at the control of the researcher. Detailed questions can be addressed, such as the utility of hydrophobic patterning or of predetermining particular sites for amino acid variation. Theory and simulation will continue to aid the design and interpretation of combinatorial experiments. Such methods will also facilitate the exploration of what is possible with the amino acids: how diverse is the set of all possible sequences that fold to a particular structure and what structures not yet seen in nature can be crafted with the amino acids? Such methods will perhaps have an even more profound impact on designing nonbiological foldamers [57••], structures about which we have much less empirical information than we do about biopolymers.

Acknowledgements

The author acknowledges support from the National Science Foundation (CHE 98-16497 and CHE 99-84752). JGS is a Cottrell Scholar of Research Corporation and is an Arnold and Mabel Beckman Foundation Young Investigator.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Kraemer-Pecore CM, Wollacott AM, Desjarlais JR: **Computational •• protein design.** *Curr Opin Chem Biol* 2001, 5:690-695.
This is a compact but excellent review on recent progress in computational methods for protein design. The authors also discuss recent efforts in designing the WW domain, a small β protein.

2. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF: **Protein design: a hierarchic approach.** *Science* 1995, 270:935-941.

3. Dunbrack R: **Rotamer libraries.** *Curr Opin Struct Biol* 2002, 12:in press.

4. Shakhnovich EI, Gutin AM: **A new approach to the design of stable proteins.** *Protein Eng* 1993, 6:793-800.

5. Jones DT: ***De novo* protein design using pairwise potentials and a genetic algorithm.** *Protein Sci* 1994, 3:567-574.

6. Onuchic JN, Luthey-Schulten Z, Wolynes PG: **Theory of protein folding: the energy landscape perspective.** *Annu Rev Phys Chem* 1997, 48:545-600.

7. Hellinga HW, Richards FM: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, 91:5803-5807.

8. Desjarlais JR, Handel TM: ***De-novo* design of the hydrophobic cores of proteins.** *Protein Sci* 1995, 4:2006-2018.

9. Johnson EC, Lazar GA, Desjarlais JR, Handel TM: **Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin.** *Structure* 1999, 7:967-976.

10. Jiang X, Farid H, Pistor E, Farid RS: **A new approach to the design • of uniquely folded thermally stable proteins.** *Protein Sci* 2000, 9:403-416.
The authors use a novel scoring function for the design of hydrophobic interiors. In addition to steric interactions, the function includes parameterizations of changes in the heat capacity and the conformational entropy upon folding. Simulated annealing was used to optimize the score. The backbone and exterior residue identities were constrained. In tests on two small proteins, in which 10–11 interior residues were varied, the native sequence was regenerated, as well as the sequences of known stable variants. Interestingly, previously designed sequences with low stability and weak cooperativity were not identified. In larger proteins tested, in which 32 and 63 residues were varied, sequence/rotamer combinations close to native were identified.

11. Jiang X, Bishop EJ, Farid RS: **A *de novo* designed protein with properties that characterize natural hyperthermophilic proteins.** *J Am Chem Soc* 1997, 119:838-839.

12. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF: **From coiled coils to small globular proteins: design of a native-like three-helix bundle.** *Protein Sci* 1998, 7:1404-1414.

13. Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF: **Solution structure and dynamics of a *de novo* designed three-helix bundle protein.** *Proc Natl Acad Sci USA* 1999, 96:5486-5491.

14. Voigt CA, Gordon DB, Mayo SL: **Trading accuracy for speed: • a quantitative comparison of search algorithms in protein sequence design.** *J Mol Biol* 2000, 299:789-803.
The authors carefully compared different methods for sequence design, including simulated annealing, genetic algorithms, mean field methods and dead end elimination (DEE). DEE most reliably finds global minima, but the authors also note that the method may be limited to 30 amino acid sites for which full amino acid variability is permitted. The authors extrapolate the results to regimes to which DEE cannot be applied. They find that both mean field and annealing approaches perform best with core residues and less reliably with residues that are fully or partially solvent exposed.

15. Gordon DB, Mayo SL: **Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem.** *J Comput Chem* 1998, 19:1505-1514.

16. Gordon DB, Mayo SL: **Branch-and-terminate: a combinatorial optimization algorithm for protein design.** *Structure* 1999, 7:1089-1098.

17. Pierce NA, Spriet JA, Desmet J, Mayo SL: **Conformational splitting: a more powerful criterion for dead-end elimination.** *J Comput Chem* 2000, 21:999-1009.

18. Dahiyat BI, Mayo SL: ***De novo* protein design: fully automated sequence selection.** *Science* 1997, 278:82-87.

19. • Marshall SA, Mayo SL: **Achieving stability and conformational specificity in designed proteins via binary patterning.** *J Mol Biol* 2001, 305:619-631.
The authors combined binary patterning with atomistic sidechain interactions to identify folding sequences of a homeodomain. Preliminary calculations were done using 'generic amino acids' to identify those sites that are most appropriate for polar or hydrophobic residues. Subject to this binary patterning, a directed search was then performed using dead end elimination. Interestingly, the authors identified an optimal binary patterning, whereby adding or subtracting hydrophobic residues adversely affects folding to stable monomers.

20. Malakauskas SM, Mayo SL: **Design, structure, and stability of a hyperthermophilic protein variant.** *Nat Struct Biol* 1998, 5:470-475.

21. Strop P, Mayo SL: **Rubredoxin variant folds without iron.** *J Am Chem Soc* 1999, 121:2341-2345.

22. Shimaoka M, Shifman JM, Jing H, Takagi L, Mayo SL, Springer TA: **Computational design of an integrin I domain stabilized in the open high affinity conformation.** *Nat Struct Biol* 2000, 7:674-678.

23. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A: ***De novo* design and structural characterization of proteins and metalloproteins.** *Annu Rev Biochem* 1999, 68:779-819.

24. • Bolon DN, Mayo SL: **Enzyme-like proteins by computational design.** *Proc Natl Acad Sci USA* 2001, 98:14274-14279.
The authors designed non-native enzymatic activity into a thioredoxin fold. The authors computationally identified promising active sites on the scaffold. The sequence was designed to stabilize the transition state of a hydrolysis reaction. The enzymes so designed had activity well above background.

25. Street AG, Mayo SL: **Computational protein design.** *Structure* 1999, 7:R105-R109.

26. • Saven JG: **Designing protein energy landscapes.** *Chem Rev* 2001, 101:3113-3130.
The author reviews recent progress in protein design from the perspective of the energy landscape theory of folding. In the context of theory, models and real systems, different issues involved in design are discussed, including target structures, energy functions, foldability criteria, search methods and the size of the amino acid alphabet.

27. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A: **ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0.** *Nucleic Acids Res* 2002, 30:301-302.

28. • Keefe AD, Szostak JW: **Functional proteins from a random-sequence library.** *Nature* 2001, 410:715-718.
A fascinating study on a random search for 'function' among random amino acid sequences. Using combinatorial methods the authors have pioneered, ATP-binding proteins were selected from a library of 10^{12} sequences.

29. Rojas NRL, Kamtekar S, Simons CT, Mclean JE, Vogel KM, Spiro TG, Farid RS, Hecht MH: ***De novo* heme proteins from designed combinatorial libraries.** *Protein Sci* 1997, 6:2512-2524.

30. Roy S, Ratnaswamy G, Boice JA, Fairman R, McLendon G, Hecht MH: **A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties.** *J Am Chem Soc* 1997, 119:5302-5306.

31. Roy S, Helmer KJ, Hecht MH: **Detecting native-like properties in combinatorial libraries of *de novo* proteins.** *Fold Des* 1997, 2:89-92.

32. Finucane MD, Tuna M, Lees JH, Woolfson DN: **Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries.** *Biochemistry* 1999, 38:11604-11612.

33. Xu GF, Wang WX, Groves JT, Hecht MH: **Self-assembled monolayers from a designed combinatorial library of *de novo* beta-sheet proteins.** *Proc Natl Acad Sci USA* 2001, 98:3652-3657.

34. Case MA, McLendon GL: **A virtual library approach to investigate protein folding and internal packing.** *J Am Chem Soc* 2000, 122:8089-8090.

35. Arndt KM, Pelletier JN, Muller KM, Alber T, Michnick SW, Pluckthun A: **A heterodimeric coiled-coil peptide pair selected *in vivo* from a designed library-versus-library ensemble.** *J Mol Biol* 2000, 295:627-639.

36. Zhao HM, Arnold FH: **Combinatorial protein design: strategies for screening protein libraries.** *Curr Opin Struct Biol* 1997, 7:480-485.

37. Giver L, Arnold FH: **Combinatorial protein design by *in vitro* recombination.** *Curr Opin Chem Biol* 1998, 2:335-338.

38. • Hoess RH: **Protein design and phage display.** *Chem Rev* 2001, 101:3205-3218.
A comprehensive review of a commonly used method to generate and display combinatorial libraries of proteins and peptides.

39. •• Moffet DA, Hecht MH: ***De novo* proteins from combinatorial libraries.** *Chem Rev* 2001, 101:3191-3203.
A review of recent work on the *de novo* combinatorial design of proteins, focusing primarily on the pioneering work of the Hecht group.

40. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH: **Protein design by binary patterning of polar and nonpolar amino-acids.** *Science* 1993, 262:1680-1685.

41. Roy S, Hecht MH: **Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids.** *Biochemistry* 2000, 39:4603-4607.

42. • Moffet DA, Case MA, House JC, Vogel K, Williams RD, Spiro TG, McLendon GL, Hecht MH: **Carbon monoxide binding by *de novo* heme proteins derived from designed combinatorial libraries.** *J Am Chem Soc* 2001, 123:2109-2115.
Heme-assisted binding of a diatomic ligand turns out to be easier to find than expected within a library of sequences patterned to form a four-helix bundle. Eight combinatorially selected heme-binding sequences bind carbon monoxide with an affinity similar to that of myoglobin. The binding properties of the proteins aren't nearly as diverse as those seen among natural heme proteins, but these *de novo* sequences serve as a useful 'reference'.

43. Moffet DA, Certain LK, Smith AJ, Kessel AJ, Beckwith KA, Hecht MH: **Peroxidase activity in heme proteins derived from a designed combinatorial library.** *J Am Chem Soc* 2000, 122:7612-7613.

44. West MW, Beasley JR, Hecht MH: **Collections of *de novo* beta-sheet proteins designed by binary patterning of polar and nonpolar amino acids.** *Protein Eng* 1997, 10:38-38.

45. West MW, Wang WX, Patterson J, Mancias JD, Beasley JR, Hecht MH: ***De novo* amyloid proteins from designed combinatorial libraries.** *Proc Natl Acad Sci USA* 1999, 96:11211-11216.

46. •• Wang WX, Hecht MH: **Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins.** *Proc Natl Acad Sci USA* 2002, 99:2760-2765.
Previously, the authors had used hydrophobic patterning consistent with β sheets that intermolecularly align 'edge on' and found these sequences did indeed form the amyloid fibrils that were expected. In this paper, they break up these edge-on interactions with a hydrophilic residue (lysine) at each edge of the β sheet. These sequences are indeed monomeric and appear to be well structured according to CD and NMR peak dispersion. These are the first examples of combinatorial β-protein design.

47. Saven JG, Wolynes PG: **Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules.** *J Phys Chem B* 1997, 101:8375-8389.

48. •• Zou JM, Saven JG: **Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure.** *J Mol Biol* 2000, 296:281-294.
The authors extend their statistical theory of sequence libraries to include negative design. The sequence space is resolved in multiple dimensions and the number of sequences is characterized according to the folded state energy and stability gap (the difference in energy between the folded state and an ensemble of unfolded conformations). Excellent agreement is observed between theoretical and exact lattice model results for both the numbers of sequences and the monomer probabilities.

49. •• Kono H, Saven JG: **Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure.** *J Mol Biol* 2001, 306:607-628.
A statistical theory of combinatorial libraries developed for combinatorial experiments. The authors used an atom-based potential and rotamer states to identify the sequence probabilities consistent with a particular structure. An effective one-body energy was introduced that relates the hydrophobicity

or solvent-exposure propensity to local β-carbon density. The calculations give good results with regard to sidechain modeling. Calculations were done that are consistent with recent combinatorial experiments on protein L. Generally, the calculations are in good agreement with the observed amino acid frequencies, despite the sampling issues that are always a concern with these comparisons. (Only 20–40 sequences were sequenced in the experiments.)

50. Kim DE, Gu HD, Baker D: **The sequences of small proteins are not extensively optimized for rapid folding by natural selection.** *Proc Natl Acad Sci USA* 1998, **95**:4982-4986.

51. Gu H, Doshi N, Kim DE, Simons KT, Santiago JV, Nauli S, Baker D: **Robustness of protein folding kinetics to surface hydrophobic substitutions.** *Protein Sci* 1999, **8**:2734-2741.

52. Koehl P, Delarue M: **Mean-field minimization methods for biological macromolecules.** *Curr Opin Struct Biol* 1996, **6**:222-226.

53. Dokholyan NV, Shakhnovich EI: **Understanding hierarchical protein evolution from first principles.** *J Mol Biol* 2001, **312**:289-307.

54. •• Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computational method to reduce the search space for directed protein evolution.** *Proc Natl Acad Sci USA* 2001, **98**:3778-3783.
The authors used a mean field theory to determine each residue's structural tolerance to mutations. This tolerance is quantified by the residue's local sequence entropy, which is a measure of the effective number of amino acids that are structurally permitted at that site. For an *in vitro* directed evolution experiment, the authors suggest that mutations that enhance stability or activity are most likely to accumulate in these high entropy regions. Multiple compensating mutations are rare in such experiments, so mutations are most likely at sites that tolerate multiple amino acids. Calculations involving subtilisin E and T4 lysozyme are consistent with the mutations observed in directed evolution experiments.

55. Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computationally focusing the directed evolution of proteins.** *J Cell Biochem* 2001:58-63.

56. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**:56-68.

57. •• Hill DJ, Mio MJ, Prince RB, Hughes TS, Moore JS: **A field guide to foldamers.** *Chem Rev* 2001, **101**:3893-4011.
A comprehensive review of nonbiological folding molecules.

EXHIBIT F

BIOCHEMISTRY

DONALD VOET
University of Pennsylvania

JUDITH G. VOET
Swarthmore College

Illustrators:
IRVING GEIS
JOHN AND BETTE WOOLSEY
PATRICK LANE

WILEY
JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

To:

Our parents, who encouraged us,
Our teachers, who enabled us, and
Our children, who put up with us.

Cover Art: One of a series of color studies of horse heart cytochrome *c* designed to show the influence of amino acid side chains on the protein's three-dimensional folding pattern. We have selected this study to symbolize the discipline of biochemistry: Both are beautiful but still in process and hence have numerous "rough edges." Drawing by Irving Geis in collaboration with Richard E. Dickerson.

Cover and part opening illustrations copyrighted by Irving Geis.

Cover Designer: Madelyn Lesure

Photo Research: John Schultz, Eloise Marion

Photo Research Manager: Stella Kupferberg

Illustration Coordinator: Edward Starr

Copy Editor: Jeannette Stiefel

Production Manager: Lucille Buonocore

Senior Production Supervisor: Linda Muriello

Library of Congress Cataloging in Publication Data:
Voet, Donald.
Biochemistry / by Donald Voet and Judith G. Voet.
p. cm.
Includes bibliographical references.
ISBN 0-471-61769-5
1. Biochemistry. I. Voet, Judith G. II. Title.
QP514.2.V64 1990
574.19'2—dc20 89-16727
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2

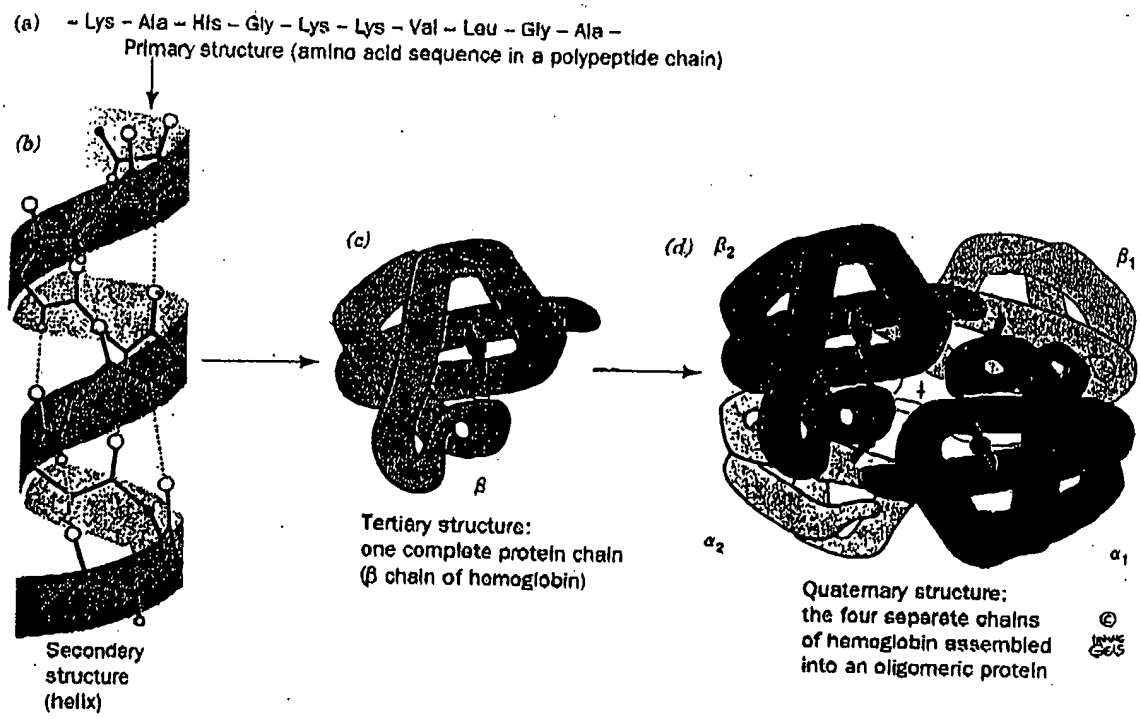


Figure 6-1
The structural hierarchy in proteins: (*a*) primary structure, (*b*) secondary structure, (*c*) tertiary structure, and (*d*) quaternary structure. [Figure copyrighted © by Irving Geis.]

Protein function can only be understood in terms of protein structure, that is, the three-dimensional relationships between a protein's component atoms. The structural descriptions of proteins, as well as those of other polymeric materials, have been traditionally described in terms of four levels of organization (Fig. 6-1):

1. A protein's **primary structure (1° structure)** is the amino acid sequence of its polypeptide chain(s).
2. **Secondary (2°) structure** is the local spatial arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains.
3. **Tertiary (3°) structure** refers to the three-dimensional structure of an entire polypeptide. The distinction between secondary and tertiary structures is, of necessity, somewhat vague; in practice, the term secondary structure alludes to easily characterized structural entities such as helices.
4. Many proteins are composed of two or more polypeptide chains, loosely referred to as **subunits**, which associate through noncovalent interactions and, in some cases, disulfide bonds. A protein's **quaternary (4°) structure** refers to the spatial arrangement of its subunits.

In this, the first of four chapters on protein structure, we discuss the 1° structures of proteins: How they are elucidated and their biological and evolutionary significance. We also survey methods of chemically synthesizing polypeptide chains. The 2°, 3°, and 4° structures of proteins which, as we shall see, are a consequence of their 1° structures, are treated in Chapter 7. In Chapter 8 we take up protein folding, dynamics, and structural evolution, and in Chapter 9 we analyze hemoglobin as a paradigm of protein structure and function.

1. PRIMARY STRUCTURE DETERMINATION

The first determination of the complete amino acid sequence of a protein, that of the bovine polypeptide hormone **insulin** by Frederick Sanger in 1953, was of enormous biochemical significance in that it definitively established that proteins have unique covalent structures. Since that time, the amino acid sequences of several thousand proteins have been elucidated. This extensive information has been of central importance in the formulation of modern concepts of biochemistry for several reasons: